Title
One and Two Locus Likelihoods Under Complex Demography

Author
Kamm, John Arthur

Publication Date
2015

Peer reviewed|Thesis/dissertation

One and Two Locus Likelihoods Under Complex Demography

by

John Arthur Kamm

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Rasmus Nielsen
Professor Montgomery W. Slatkin

Fall 2015

One and Two Locus Likelihoods Under Complex Demography

Abstract

One and Two Locus Likelihoods Under Complex Demography

by

John Arthur Kamm

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Yun S. Song, Chair

The coalescent is a random process that describes the genealogy relating a sample of individuals, and provides a probability model that can be used for likelihood-based inference on genetics data. For example, coalescent models may include recombination, natural selection, population size crashes and growth, and migrations, and thus can be used to learn the strength of these biological and demographic forces. Unfortunately, computing the likelihood of data remains a challenging problem in many of these coalescent models.

In this dissertation, I develop new equations and algorithms for computing coalescent likelihoods at one or two loci, and apply them to inference problems in a composite likelihood framework. I begin by developing an algorithm for the one-locus case, computing the *site frequency spectrum* (the distribution of mutant allele counts) under complex demographic histories with population size changes (including exponential growth), population splits, population mergers, and admixture events. This method improves on the runtime and numerical stability of previous approaches, and can successfully infer demographic histories that would otherwise be too computationally challenging to consider. I then consider the two-locus case, and derive a formula for the likelihood at a pair of sites under a variable population size history; this formula scales to tens of individuals. In addition to this exact formula, I also develop a highly efficient importance sampler to compute the same likelihood. I apply these results to the problem of inferring recombination rates under variable population size.

To Joy, and to my parents.

Contents

List of Figures

List of Tables

Acknowledgments

My advisor Yun S. Song has been a fantastic guide and mentor, and I want to thank him for playing such a huge role in my growth as a researcher. I have particularly enjoyed our many stimulating discussions over the years, and I have benefitted greatly from his insight, wisdom, and enthusiasm for science.

This dissertation includes work that was done with a number of collaborators. I want to thank Yun, Jonathan Terhorst, Jeff Spence, and Jeff Chan for their contributions. I also want to thank Matthias Steinrücken, Anand Bhaskar, Sara Sheehan, Kelley Harris, and Emilia Wieczorek for other fruitful collaborations during my Ph.D. career. In addition, I am thankful to members of Yun's group, members of the Center for Theoretical and Evolutionary Genomics, fellow Statistics students, and my former housemate Guy Isely, for many stimulating conversations on academic and unrelated topics.

I am grateful to the members of my dissertation and qualifying exam committees – Yun, Rasmus Nielsen, Monty Slatkin, and Steve Evans. Rasmus and Monty in particular have provided much useful feedback and interesting discussion throughout my Ph.D., especially through our interactions at the Center for Theoretical and Evolutionary Genomics. I am also grateful to David Brillinger, who I was a graduate student instructor for, and from whom I learned a great deal about statistics and teaching. Bin Yu, Cari Kaufman, Elizabeth Purdom, and Michael Jordan are other faculty whose excellent classes have had a big impact on my statistical thinking. In addition, several professors from my Stanford undergrad, especially Mehran Sahami, have played an important role in my development as a teacher and researcher.

On a more personal note, I want to thank my parents and my wife Joy. I would not be here without my parents' unconditional support, and their encouragement of my educational pursuits throughout my life. It was nice to have them close to Berkeley, and I was happy for the many warm meals at home during my Ph.D. Finally, I am extremely lucky to be married to Joy. Her constant encouragement and support have kept me going through the many ups and downs of my Ph.D., while her common sense and hilarious, crass sense of humor have always put things in perspective and kept me grounded.

Chapter 1

Introduction

The *coalescent* (Kingman, 1982a,b,c) is a stochastic process commonly used to model data in population genetics. Coalescent models may include selection, recombination, and complex demographic history, and thus provide a framework for inferring these evolutionary forces. Unfortunately, computing likelihoods under the coalescent is a difficult problem, especially for large, modern genomic datasets.

If the full likelihood is intractable, one strategy is to use a *composite likelihood* method, by computing likelihoods of small subsets of the data and combining these in some way. Still, challenges remain in computing sampling probabilities at even 1 or 2 sites of the genome, especially under complex demographic scenarios. At a single site, methods exist for computing the sample frequency spectrum (SFS), or distribution of allele counts (Gutenkunst et al., 2009; Chen, 2012); however, these methods have trouble scaling to a large number of individuals or populations. At a pair of sites, there are no existing methods for computing sampling probabilities under a changing population size.

In this dissertation, we address these issues with new formulas and algorithms. We consider the one locus case in Chapters 2 and 3, and the two locus case in Chapter 4. In Chapter 2, we introduce mathematical and computational results that improve the speed and numerical stability of computing the SFS, but mostly focus on the case without admixture. In Chapter 3, we focus on the case with admixture, generalizing some of the improvements from Chapter 2. We also show how to compute expectations of a wide range of genealogical and summary statistics, and we apply a composite likelihood to estimate a complex 6-population history with admixture and exponential growth. In Chapter 4, we show how to compute the two-locus sampling probability under a changing but piecewise constant population size. We introduce two distinct methods, an exact formula and a highly efficient importance sampler. We show how accounting for population size changes improves the inferred recombination map under a two-locus composite likelihod.

This dissertation focuses on the neutral model without selection. In Chapter 5, we conclude by considering some future directions and open problems, such as generalizing our results to include natural selection.

1.1 Preliminaries

We begin by reviewing the neutral coalescent at a single locus under a constant, panmictic population size.

1.1.1 The coalescent

The simplest population genetics model is the one-locus, panmictic *Wright-Fisher* process. In this model, a population consists of N individual alleles, and each generation is obtained from the previous generation by sampling with replacement. That is, every individual allele independently "chooses" its parent in the previous generation.

If we take a sample of size n at the present, the samples will be related by a *rooted tree* or *genealogy*: tracing the samples backward in time through their ancestors, every time two lineages find a common ancestor, the total number of ancestors will drop by 1, until there is only a single *root* or *most recent common ancestor* left (with the time to most recent common ancestor $T_{\text{MRCA}} < \infty$ almost surely).

If we let $N \to \infty$, and scale time as $\frac{1}{N}$ per generation, then the genealogy converges in distribution to a stochastic process called the *coalescent*, defined as follows. At the present we have n lineages. Going backwards in time, every pair of lineages *coalesces* (finds a common ancestor) at rate 1, so that the waiting time until the next coalescent event is $\text{Exp}(\binom{m}{2})$ when there are m ancestral lineages remaining. The resulting sample genealogy is a rooted, ultrametric, binary tree.

In fact, this limit holds for other population models besides the Wright-Fisher process. This justifies using the coalescent to model genealogies under more realistic mating models. The main requirement is that the distribution of offspring per parent decays rapidly enough (Cannings, 1974). More precisely, let ν_i be the number of offspring of the ith individual at the present generation, let $c_N = \frac{\mathbb{E}[\nu_1(\nu_1-1)]}{N-1}$, and $d_N = \frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{(N-1)(N-2)}$. If $(\nu_1, \ldots, \nu_N)$ is exchangeable, and $c_N \to 0$ and $\frac{d_N}{c_N} \to 0$ as $N \to \infty$, then the sample genealogy converges in distribution to the coalescent (scaling time so that each generation has length c_N).

1.1.2 Sampling probability and frequency spectrum

Now suppose that mutations arise on the coalescent tree as a Poisson point process with rate $\frac{\theta}{2}$. These mutations cause the samples to have different allelic types. For each allelic type k, let n_k be the number of copies of k in the sample, so $\sum_k n_k = n$.

We are interested in computing the *sampling probability* $\mathbb{P}(\{n_k\})$, which we review here for three different cases: infinite alleles, finite alleles, and at an infinitessimal site.

In all three cases, it is useful to consider *identity by descent* (IBD), which can be modeled by a *Chinese restaurant process* (Aldous, 1985; Durrett, 2008). Say 2 points on the coalescent tree are IBD if the path between them contains no mutations. Consider when there were m lineages in IBD with the present sample, for $0 \le m \le n-1$. Imagine these as m "customers" in a restaurant, each at a "table" corresponding to its allelic type. Going forward in time,

eventually there are $m+1$ lineages in IBD with the sample; an $(m+1)$th customer walks in, and is seated at a new table with probability $\frac{\theta}{\theta+m}$, and at an existing table with j customers with probability $\frac{j}{\theta+m}$. This is because, going backwards in time from the $m+1$ lineages, IBD decreases due to mutation at rate $\frac{\theta(m+1)}{2}$, and decreases due to coalescence at rate $\frac{m(m+1)}{2}$.

Infinite alleles

In this model, each mutation gives rise to a new allelic type. For a sample with K unique alleles, label the observed alleles by a uniform permutation of $1, \ldots, K$, so $\sum_{k=1}^{K} n_k = n$. Then the Chinese restaurant process yields

$$\mathbb{P}(n_1, \ldots, n_K) = \binom{n}{n_1, \ldots, n_K} \frac{\theta^K \prod_{k=1}^{K} (n_k - 1)!}{K! \prod_{m=0}^{n-1} (\theta + m)} = \frac{\theta^K n!}{K! (\theta)_{n\uparrow} \prod_{k=1}^{K} n_k}$$

with $(a)_{b\uparrow} = \prod_{i=0}^{b-1} (a+i)$ the rising factorial. Note that *Ewens' sampling formula* (Ewens, 1972) gives a similar probability, but ignores the allele labels, and so is equal up to a combinatorial factor.

Finite alleles

In this model, there are a finite number of alleles $\{1, \ldots, K\}$, with allele j mutating to allele k at rate $\frac{\theta}{2} P_{jk}$. In general there is no closed form solution for the sampling probability. However, in the special case where $P_{jk} = P_k$ doesn't depend on j,

$$\mathbb{P}(n_1, \ldots, n_K) = \binom{n}{n_1, \ldots, n_K} \frac{\prod_k (\theta P_k)_{n_k\uparrow}}{(\theta)_{n\uparrow}}$$

which follows from noting the Chinese restaurant process yields a Dirichlet-multinomial distribution, if a table has allele k with probability P_k.

Infinitessimal site

Typical DNA sequences are made up of a large number of positions or *sites*, each with a very small mutation rate. This is approximated by the *infinite sites* model, where each individual allele is composed of an infinite number of sites, each with an infinitesimally small mutation rate, so that the total mutation rate over all sites is $\frac{\theta_{\text{tot}}}{2} \in (0, \infty)$.

The *sample frequency spectrum* (SFS) ξ_k is the expected branch length with k leafs, $1 \leq k < n$. Since mutations hit the genealogy with rate $\frac{\theta_{\text{tot}}}{2}$, the expected number of sites with k derived (mutant) alleles in the sample is $\frac{\theta_{\text{tot}}}{2} \xi_k$.

For the standard coalescent, $\xi_k = \frac{2}{k}$. To see this, consider a site with very small mutation rate $\theta/2 \ll 1$. Let n_0 be the number of copies with the oldest observed allele, and $x = n - n_0$ the count of more recent alleles. From the Chinese restaurant process,

$$\mathbb{P}_\theta(x = k) = \binom{n-1}{n-k-1} \frac{(\theta)_{k\uparrow} (n-k-1)!}{(\theta+1)_{n-1\uparrow}} = \frac{\theta}{k} + O(\theta^2) \tag{1.1}$$

because $n-k-1$ out of the last $n-1$ customers sit at the first table. Returning to the infinite sites scenario, we can obtain ξ_k by applying (1.1) at each site. In particular, we consider L sites, each with mutation rate $\frac{\theta_{\text{tot}}}{2L}$, and send $L \to \infty$ to obtain the infinite sites model. Then the expected number of sites with k derived alleles is

$$\begin{aligned}\frac{\theta_{\text{tot}}}{2}\xi_k &= \lim_{L\to\infty} L\mathbb{P}_{\theta_{\text{tot}}/L}(x=k) \\ &= \lim_{L\to\infty} L\left(\frac{\theta_{\text{tot}}/L}{k} + O\left(\frac{\theta_{\text{tot}}^2}{L^2}\right)\right) = \frac{\theta_{\text{tot}}}{k}\end{aligned}$$

and so $\xi_k = \frac{2}{k}$.

Note the SFS is related to sampling probabilities in two ways. First, the sampling probability of derived alleles given a mutation on the tree is

$$\mathbb{P}(x = k \mid \text{mutation}) = \frac{\xi_k}{\sum_{j=1}^{n-1}\xi_j} = \frac{1}{kH_{n-1}}$$

with $H_m = \sum_{i=m}^{m}\frac{1}{i}$ the mth harmonic number. Secondly, from (1.1) and $\xi_k = \frac{2}{k}$, we have $\mathbb{P}_\theta(x=k) = \frac{\theta}{2}\xi_k + o(\theta)$, i.e. ξ_k provides a first-order approximation to the sampling probability under small mutation rate. In other words, the SFS is

$$\xi_k = 2\frac{d}{d\theta}\mathbb{P}_\theta(x=k)\Big|_{\theta=0}$$

the derivative around $\frac{\theta}{2} = 0$, of the sampling probability for k derived copies and $n-k$ ancestral copies.

1.1.3 Moran model

The Moran model is a continuous-time population model with sample genealogies exactly equal to the coalescent (Moran, 1958; Ethier and Kurtz, 1993; Donnelly and Kurtz, 1999). We will make extensive use of this equivalence throughout the thesis.

The Moran model is a forward-in-time process, defined as follows. At each time there are n individual alleles, which mutate at rate $\frac{\theta}{2}$, as before. In addition, *copying events* occur, where an individual copies its allele onto some other individual, replacing the existing allele. This happens at rate $\frac{1}{2}$ for each pair of individuals and each direction of copying. The coalescent is embedded within the Moran model because, tracing the ancestry of a sample backwards in time, each copying event is like a coalescence, and coalescence occurs at rate 1 per pair of lineages.

Note the Moran model is usually scaled differently: in particular, the Moran model is typically defined to have N lineages, with copying happening at rate $\frac{1}{2N}$ instead of $\frac{1}{2}$. We avoid this convention, so that we can directly apply the equivalence of the Moran model and the coalescent, without having to first scale time by a factor of $\frac{1}{N}$.

1.2 Overview

We now outline the remainder of this dissertation. Chapter 2 is based on a preprint by Kamm, Terhorst, and Song (2015b), while Chapter 4 is based on a preprint by Kamm, Spence, Chan, and Song (2015a).

1.2.1 Efficient computation of the multipopulation SFS

In Chapter 2, we consider computing the joint SFS for $\mathcal{D}$ sampled populations that may change over time. The joint SFS is the expected $\mathcal{D}$-dimensional histogram of derived allele counts, per unit mutation rate. As before, we may equivalently define the SFS as an expected branch length, or as the derivative of a sampling probability around $\frac{\theta}{2} = 0$ (Griffiths and Tavaré, 1998).

There has been much interest in analyzing joint SFS data from multiple populations to infer parameters of complex demographic histories, including variable population sizes, population split times, migration rates, admixture proportions, and so on. This requires accurate computation of the SFS under a given demographic model. Although much methodological progress has been made, existing methods suffer from numerical instability and high computational complexity when multiple populations are involved and the sample size is large. We present new analytic formulas and algorithms that enable oaccurate, efficient computation of the joint SFS for thousands of individuals sampled from hundreds of populations related by a complex demographic model with arbitrary population size histories (including piecewise-exponential growth). Through an empirical study we demonstrate the improvements to numerical stability and computational complexity.

The main algorithm in Chapter 2 is based on a previous method that implicitly integrates over coalescent trees (Chen, 2012), but with two innovations that substantially improve its stability and speed:

1. We apply formulas from Polanski and Kimmel (2003) to efficiently and stably compute certain terms, which we call the *truncated SFS*, and which correspond to the frequency of mutations arising in each part of the demographic history.

2. We replace the coalescent within each population by an equivalent Moran model. This yields a substantial speedup because the Moran model has fewer states to integrate over than the coalescent.

Note that initially, we only apply the Moran model speedup to demographies without admixture. In the following Chapter 3, we generalize the Moran model speedup to handle admixture as well.

1.2.2 The Moran model for the SFS with admixture

In Chapter 3, we continue our discussion of the multipopulation SFS for complex demography. In particular, we extend the Moran model speedup from Chapter 3 to handle discrete admixture events. To do this, we construct a *lookdown model* of the multipopulation Moran process (Donnelly and Kurtz, 1996), which is a version of the Moran model with a countably infinite number of lineages. By embedding this lookdown model within the demographic history, we are able to construct a more efficient algorithm for the SFS under admixture. This algorithm is purely based on the Moran model, and does not use the coalescent directly.

In addition, we note that our algorithm can efficiently compute a number of *linear summary statistics* of the SFS. These statistics include $\mathbb{E}[T_{\text{MRCA}}]$, $\mathbb{E}[\text{Total Branch Length}]$, and other classical statistics from population genetics.

Finally, we examine the problem of inferring complex demographic history using the joint SFS. We consider a composite likelihood approach, searching for the maximum composite likelihood estimate using gradient descent and automatic differentiation. Using simulations, we show that we can quickly and accurately infer a 6 population demography with 18 parameters, including admixture and exponential growth.

1.2.3 Two loci under changing population size

In Chapter 4, we turn our attention from the one-locus SFS to the sampling probability at two linked sites. Two-locus sampling probabilities have played a central role in devising an efficient composite likelihood method for estimating fine-scale recombination rates. Due to mathematical and computational challenges, these sampling probabilities are typically computed under the unrealistic assumption of a constant population size, and simulation studies have shown that resulting recombination rate estimates can be severely biased in certain cases of historical population size changes. To alleviate this problem, we develop two distinct methods to compute the sampling probability for variable population size functions that are piecewise constant. The first is a novel formula that can be evaluated by numerically exponentiating a large but sparse matrix. The second method is importance sampling on genealogies, based on a characterization of the optimal proposal distribution that extends previous results to the variable-size setting. The resulting proposal distribution is highly efficient, with an average effective sample size (ESS) of nearly 98% per sample. Using our methods, we study how a sharp population bottleneck followed by rapid growth affects the correlation between partially linked sites. Then, through an extensive simulation study, we show that accounting for population size changes under such a demographic model leads to statistically significant and in some cases dramatic improvements in recombination rate estimation.

Chapter 2

Truncated and multipopulation SFS

In this chapter, we present a method for computing the expected joint sample frequency spectrum (SFS), a summary statistic on DNA sequences which lies at the core of a large number of empirical investigations and inference procedures in population genetics (Wakeley and Hey, 1997; Griffiths and Tavaré, 1998; Nielsen, 2000; Gutenkunst et al., 2009; Coventry et al., 2010; Gazave et al., 2014; Gravel et al., 2011; Nelson et al., 2012; Excoffier et al., 2013; Jenkins et al., 2014; Bhaskar et al., 2015). The joint SFS is of interest because it maps complex demographic models involving population size changes, population splits, migration, and admixture to a low-dimensional vector, thus providing a useful analytic tool for performing inference when confronted with a large number of sampled DNA sequences.

We argue both theoretically and by simulation that existing methods in population genetics cannot scale to the problem sizes encountered in modern genetic analyses. Our primary contribution is a novel algorithm (and accompanying open-source software package) which is several orders of magnitude faster (in the number of sampled individuals n) than existing approaches. Moreover, by careful algorithmic design we are able to mitigate certain numerical issues (e.g., underflow and catastrophic cancellation) that are commonly encountered in our problem setting, but for which little work has been done previously. The combined effect of these innovations is to permit the analysis of much larger data sets, which will lead to improved inference in population genetics.

The rest of the chapter is organized as follows: In Section 2.1, we survey related work and summarize our main results. Section 2.2 presents the theoretical results that lead to the improved algorithm described in Section 2.3. Runtime and numerical accuracy results are discussed in Section 2.4. Mathematical proofs of our theoretical results are deferred to Section 2.5.

Software availability: The algorithms presented in this chapter are implemented in a software package called *momi* (MOran Models for Inference), which is freely available at `https://github.com/jackkamm/momi`.

2.1 Background and summary

2.1.1 Existing work

Two approaches exist for computing the joint SFS (the multidimensional histogram of mutant allele counts). One is based on the Wright-Fisher diffusion (Ewens, 2004), a stochastic process which describes the frequency trajectory of a mutant allele segregating in a population as time moves forward. The diffusion framework has the advantage of being applicable to arbitrary demographic models, but its computational complexity grows exponentially with the number of populations. Also, it requires numerically solving a system of partial differential equations, which can be difficult in practice. For these reasons, current implementations (Gutenkunst et al., 2009; Gravel et al., 2011; Lukić and Hey, 2012) of the diffusion approach are limited to analyzing no more than four populations at once.

An alternative class of methods, which includes ours, relies on the coalescent, which is dual to the Wright-Fisher diffusion. Using the coalescent, one computes the expected SFS by integrating over all genealogies underlying the sample. This can be done either via Monte Carlo or analytically. Monte Carlo integration (Nielsen, 2000) can effectively handle arbitrary demographic histories with a large number of populations, and Excoffier et al. (2013) have recently developed a useful implementation. However, when the number $\mathcal{D}$ of populations (or demes) is moderate to large, most of the $O(n^{\mathcal{D}})$ SFS entries will be unobserved in simulations, and thus the Monte Carlo integral may naively assign a probability of 0 to observed SNPs. Monte Carlo computation of the expected SFS thus requires careful regularization techniques to avoid degeneracy.

An alternative to the Monte Carlo approach is to compute the expected SFS exactly via analytic integration over coalescent genealogies (Wakeley and Hey, 1997; Griffiths and Tavaré, 1998). For a demography involving multiple populations, this can be done by a dynamic program (Chen, 2012, 2013). This algorithm is more complicated and less general than both the Monte Carlo and diffusion approaches: while it can handle population splits, merges, size changes, and instantaneous gene flow, it is difficult to include continuous gene flow between populations. However, it scales well to a large number $\mathcal{D}$ of populations, since it only computes entries of the SFS that are observed in the data, and ignores the $O(n^{\mathcal{D}})$ SFS entries that are not observed.

2.1.2 Summary of results

Unfortunately, existing coalescent-based algorithms (Wakeley and Hey, 1997; Chen, 2012, 2013) do not scale well to large sample size n, either in terms of running time or numerical stability. These algorithms rely on large alternating sums that explode with n and exhibit catastrophic cancellation. To circumvent these problems, we obtain new results for computing the *truncated SFS*, a key quantity needed to compute the joint SFS for multiple populations. For a fixed time τ, the truncated SFS gives the expected number of mutations arising in the time interval $[0, \tau)$ that are found in $k = 1, 2, \ldots, n$ individuals sampled at

time 0. We provide an algorithm for computing this quantity efficiently and in a numerically stable manner.

For general demographic histories, the complexity of the dynamic program devised by Chen (2012, 2013) is $O(n^5V+WL)$, where V is the number of populations (vertices) throughout the history, L is the number of distinct SFS entries to be computed, and W is a term that depends on n and the graph structure of the demography; in this chapter, we improve this to $O(n^2V + WL)$. For the special case of a tree-shaped demography without migration or admixture, Chen's algorithm gives $W = O(n^4V)$. By using the *Moran model*, we further improve this to $W = O((n^2+n\mathcal{T})V)$, where $\mathcal{T}$ is the number of matrix-vector multiplications used to compute a certain matrix exponential (Section 2.3.2). Our Moran-based speedup generalizes to non-tree demographies, but we leave this generalization to Chapter 3.

We show through an empirical study that our algorithm is not only orders of magnitude faster, but also more numerically stable.

Lastly, we note that our algorithm relies on ideas similar to those found in Bryant et al. (2012) and De Maio et al. (2013), but with a different focus. Those methods aim to compute a "phylogenetic SFS", in which mutations are allowed to be recurrent and time scales are sufficiently long that population size can be assumed to be constant. In contrast, our method considers an infinite sites model (Kimura, 1969) without recurrent mutation, and can handle arbitrary population size change functions. These features make it more appropriate for use in a population genetic setting.

2.2 Theoretical results on the truncated SFS

In this section we study the truncated sample frequency spectrum, which is the key object needed to compute the joint SFS in our algorithm.

2.2.1 Notation for the coalescent and the SFS

We define the coalescent $\{\mathcal{C}_t^n\}_{t\geq 0}$ on n leaves as the backward-in-time Markov jump process whose value at time t is a partition of $\{1,\dots,n\}$, and at time t, each pairs of blocks in $\mathcal{C}_t^n$ merge with rate $\alpha(t)$. We also call $\frac{1}{\alpha(t)}$ the *population size history function*. We usually drop the dependence on n, and write $\mathcal{C}_t = \mathcal{C}_t^n$. We sometimes denote a dependence on n through the probability $\mathbb{P}_n$ and the expectation $\mathbb{E}_n$. So if $Y(\mathcal{C}^n)$ denotes a random variable of the process $\mathcal{C}^n$, we usually write $\mathbb{E}_n[Y]$ instead of $\mathbb{E}[Y(\mathcal{C}^n)]$.

A sample path of $\{\mathcal{C}_t\}_{t\geq 0}$ can be viewed as a rooted ultrametric binary tree with n leaves, labeled $1,\dots,n$, corresponding to sampled individuals. The tree extends backwards in time and each branch represents a partition block of the process such that $\mathcal{C}_t$ corresponds to the partition induced on $\{1,\dots,n\}$ by cutting the tree at height t. To generate data, we drop mutations on the tree at rate $\frac{\theta}{2}$. Let $\mathcal{X} \subset \{1,\dots,n\}$ be the subsample with at least one mutation since the common ancestor.

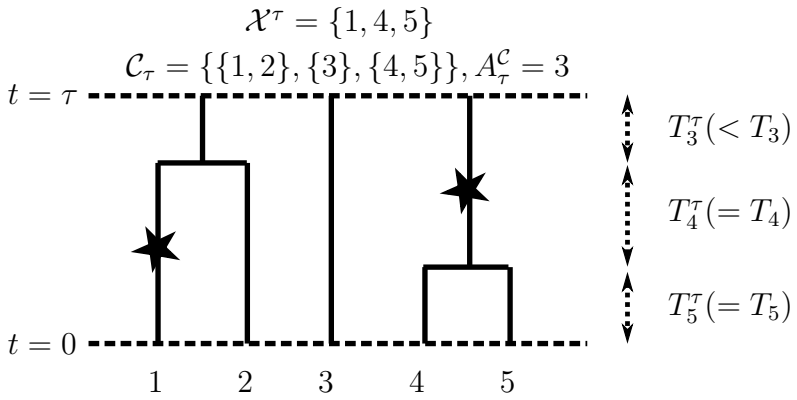


Figure 2.1: A sample path of the coalescent truncated at time τ. Star symbols denote mutations, while $\mathcal{X}^\tau$ denotes the set of leaves under those mutations. T^τ_k denotes the waiting time in the interval $[0, \tau)$ while there are k lineages.

We define the sample frequency spectrum ξ_k, for $0 < k < n$, as the first order Taylor series coefficient of $\mathbb{P}_n(|\mathcal{X}| = k)$ in the mutation rate,

$$\mathbb{P}_n(|\mathcal{X}| = k) = \frac{\theta}{2}\xi_k + o(\theta).$$

We will generally refer to ξ_k as the sample frequency spectrum (SFS). We also note two alternative definitions of the SFS. First, ξ_k is the expected number of mutations with k descendants when $\frac{\theta}{2} = 1$. Second, $\frac{1}{\binom{n}{|K|}}\xi_{|K|}$ is the expected length of the branch whose leaf set is $K \subset \{1, \ldots, n\}$. More specifically, let $\mathbb{I}$ denote the indicator function, and define $\mathcal{B}_K := \int_0^\infty \mathbb{I}_{K \in \mathcal{C}_t} dt$. Then

$$\frac{1}{\binom{n}{|K|}}\xi_{|K|} = \mathbb{E}_n[\mathcal{B}_K].$$

The equivalence of these alternate definitions follows from previous results in Griffiths and Tavaré (1998); Jenkins and Song (2011); Bhaskar et al. (2012).

We now consider truncating the coalescent with mutation at time τ, as illustrated in Figure 2.1. Let $\mathcal{X}^\tau$ denote the set of leaves under mutations occurring in the time interval $[0, \tau)$. We define the *truncated* SFS $f^\tau_n(k)$ according to

$$\mathbb{P}_n(|\mathcal{X}^\tau| = k) = \frac{\theta}{2} f^\tau_n(k) + o(\theta),$$

where $f^\tau_n(k)$ corresponds to the total expected length of all branches in the time interval $[0, \tau)$ each of which subtends k leaves. Note that $\xi_k \equiv f^\infty_n(k)$. Using the truncated SFS $f^\tau_n(k)$ for each population appearing in a demographic history, where τ denotes the length of time a particular population exists, it is possible to compute the joint SFS for multiple related populations (Chen, 2012). In Section 2.3.1, we describe a dynamic program algorithm for computing the joint SFS for multiple populations related by a complex demography, and the way in which this algorithm uses the truncated SFS $f^\tau_\nu(k)$.

2.2.2 Previous work on the truncated SFS

The key challenge is to compute the truncated SFS $f_n^\tau(k)$. The approach taken by Chen (2012) to tackle this problem is as follows. Let $A_t^{\mathcal{C}}$ be the random variable which equals the number of coalescent lineages at time t ancestral to the sample. This is a pure death process which decreases by one each time a coalescence event occurs, until finally reaching the absorbing state 1 when all individuals have found a common ancestor. More precisely, $A_t^{\mathcal{C}} = |\mathcal{C}_t|$ and the rate of transition from m to $m-1$ is given by $\lambda_{m,m-1}^{\mathcal{C}}(t) = \binom{m}{2}\alpha(t)$. We define a conditional version of the SFS via $f_n^\tau(k \mid A_\tau^{\mathcal{C}} = m)$ according to

$$\mathbb{P}_n(|\mathcal{X}^\tau| = k \mid A_\tau^{\mathcal{C}} = m) = \frac{\theta}{2} f_n^\tau(k \mid A_\tau^{\mathcal{C}} = m) + o(\theta). \tag{2.1}$$

Here, $f_n^\tau(k \mid A_\tau^{\mathcal{C}} = m)$ is the total expected length of all branches each subtending k leaves given that there are m ancestors at time τ.

Chen's approach to computing the unconditional SFS $f_\nu^\tau(k)$ is to use the decomposition

$$f_\nu^\tau(k) = \sum_{m=1}^{n-k+1} \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m) f_\nu^\tau(k \mid A_\tau^{\mathcal{C}} = m). \tag{2.2}$$

The first term in the summand, $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$, can be computed in at least three ways: by numerically exponentiating the rate matrix of $A^{\mathcal{C}}$, by computing an alternating sum with $O(\nu)$ terms (Tavaré, 1984), or by solving a recursion described in Section 2.5.1. We note that the recursion described in Section 2.5.1 has the advantage of computing all values of $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$, $m \le \nu \le n$, in $O(n^2)$ time.

The second term $f_\nu^\tau(k \mid A_\tau^{\mathcal{C}} = m)$ in the summand of (2.2) is computed in Chen (2012) as

$$f_\nu^\tau(k \mid A_\tau^{\mathcal{C}} = m) = \sum_{i=m}^{\nu} i p_{\nu,i}^{k,1} \mathbb{E}_\nu[T_i \mid A_\tau^{\mathcal{C}} = m], \tag{2.3}$$

where

$$p_{\nu,i}^{k,j} := \begin{cases} \frac{\binom{k-1}{j-1}\binom{\nu-k-1}{i-j-1}}{\binom{\nu-1}{i-1}}, & \text{if } k \ge j > 0 \text{ and } \nu - k \ge i - j > 0, \\ 1, & \text{if } j = k = 0 \text{ or } i - j = \nu - k = 0, \\ 0, & \text{else,} \end{cases}$$

is the transition probability of the Pólya urn model, starting with $i-j$ white balls and j black balls, and ending with $\nu - k$ white balls and k black balls (Johnson and Kotz, 1977), and

$$T_i := \int_0^\tau \mathbb{I}_{A_t^{\mathcal{C}} = i} dt$$

is the length of time in $[0, \tau)$ where there are i ancestral lineages to the sample, as illustrated in Figure 2.1. Chen (2012) provides a formula for the conditional expectation $\mathbb{E}_\nu[T_i \mid A_\tau^{\mathcal{C}} = m]$ for the case of constant population size, which he later extends (Chen, 2013) to the case of an exponentially growing population. However, these formulas involve a large alternating sum with $O(\nu^2)$ terms. Thus, computing $\mathbb{E}_\nu[T_i \mid A_\tau^{\mathcal{C}} = m]$ for every value of i, m, ν, as required to compute $\{f_\nu^\tau(k)\}_{k \le \nu \le n}$ with (2.2) and (2.3), takes $O(n^5)$ time with these formulas. In addition, large alternating sums are numerically unstable due to catastrophic cancellation (Higham, 2002), and so these formulas require the use of high-precision numerical libraries, further increasing runtime.

2.2.3 An efficient, numerically stable algorithm for computing the truncated SFS

Here, we present a numerically stable algorithm to compute $\{f_\nu^\tau(k) \mid 1 \le k \le \nu \le n\}$ in $O(n^2)$ time instead of $O(n^5)$ time. Our approach utilizes the following two lemmas:

Lemma 1. *The entry $f_n^\tau(n)$ of the truncated SFS is given by*

$$f_n^\tau(n) = \tau - \sum_{k=1}^{n-1} \frac{k}{n} f_n^\tau(k). \tag{2.4}$$

Lemma 2. *For all $1 \le k \le \nu$, the truncated SFS $f_\nu^\tau(k)$ satisfies the linear recurrence*

$$f_\nu^\tau(k) = \frac{\nu - k + 1}{\nu + 1} f_{\nu+1}^\tau(k) + \frac{k+1}{\nu+1} f_{\nu+1}^\tau(k+1). \tag{2.5}$$

We prove Lemma 1 in Section 2.5.2. We note here that our proof also yields the identity $\mathbb{E}[T_{\text{MRCA}}] = \sum_{k=1}^{n-1} \frac{k}{n} \xi_k$, where T_{MRCA} denotes the time to the most recent common ancestor of the sample; to our knowledge, this formula is new. A proof of Lemma 2 is provided in Section 2.5.3.

We now sketch our algorithm. For a given n, we show below that all values of $f_n^\tau(k)$, for $1 \le k < n$, can be computed in $O(n^2)$ time. We then compute $f_n^\tau(n)$ using Lemma 1 in $O(n)$ time. Finally, using $f_n^\tau(k)$ for $1 \le k \le n$ as boundary conditions, Lemma 2 allows us to compute all $f_\nu^\tau(k)$, for $\nu = n-1, n-2, \ldots, 2$ and $k = 1, \ldots, \nu$, in $O(n^2)$ time.

We now describe how to compute the aforementioned terms $f_n^\tau(k)$, for all $k < n$, in $O(n^2)$ time. We first recall the result of Polanski and Kimmel (2003) which represents the untruncated SFS ξ_k, for $1 \le k \le n-1$, as

$$\xi_k = \sum_{m=2}^{n} W_{n,k,m} c_m, \tag{2.6}$$

where

$$c_m := \mathbb{E}_m[T_m] = \int_0^\infty t\binom{m}{2}\alpha(t)\exp\left[-\binom{m}{2}\int_0^t \alpha(x)dx\right]dt$$
$$= \int_0^\infty \exp\left[-\binom{m}{2}\int_0^t \alpha(x)dx\right]dt \tag{2.7}$$

denotes the waiting time to the first coalescence for a sample of size m, and $W_{n,k,m}$ are universal constants that are efficiently computable using the following recursions (Polanski and Kimmel, 2003):

$$\begin{aligned} W_{n,k,2} &= \frac{6}{n+1}, \\ W_{n,k,3} &= 30\frac{(n-2k)}{(n+1)(n+2)}, \\ W_{n,k,m+2} &= -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)}W_{n,k,m} + \frac{(3+2m)(n-2k)}{m(n+m+1)}W_{n,k,m+1}, \end{aligned} \tag{2.8}$$

for $2 \le m \le n-2$. The key observation is to note that, in a similar vein as (2.6), we have:

Lemma 3. *The truncated SFS $f_n^\tau(k)$, for $1 \le k \le n-1$, can be written as*

$$f_n^\tau(k) = \sum_{m=2}^{n} W_{n,k,m}c_m^\tau, \tag{2.9}$$

where c_m^τ is a truncated version of (2.7)*:*

$$c_m^\tau := \mathbb{E}_m[T_m^\tau] = \int_0^\tau \exp\left[-\binom{m}{2}\int_0^t \alpha(x)dx\right]dt. \tag{2.10}$$

We prove Lemma 3 in Section 2.5.4. For piecewise-exponential $\alpha(t)$, c_m^τ can be computed explicitly using formulas from Bhaskar et al. (2015). Using (2.8), we can compute all values of $W_{n,k,m}$, for $1 \le k \le n$ and $2 \le m \le n$, in $O(n^2)$ time. Then, using (2.9), all values of $f_n^\tau(k)$, for $1 \le k \le n-1$ can be computed in $O(n^2)$ time.

Note that the above algorithm not only significantly improves computational complexity, but also resolves numerical issues, since it allows us to avoid computing the expected times $\mathbb{E}_\nu[T_i \mid A_\tau^{\mathcal{C}} = m]$, which are alternating sums of $O(n^2)$ terms and are numerically unstable to evaluate for large values of n (say, $n > 50$).

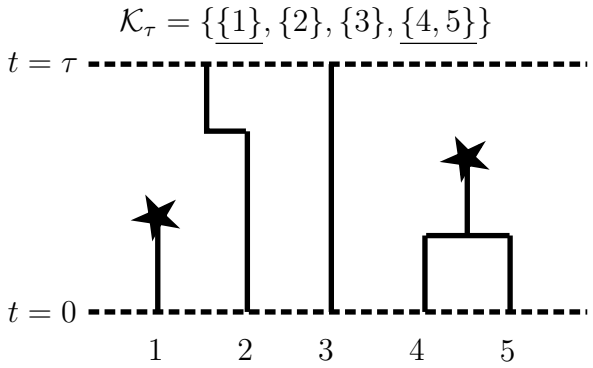


Figure 2.2: The coalescent with killing for the genealogy in Figure 2.1. Note that $\mathcal{K}_\tau$ is a marked partition, with the blocks killed by mutations in $[0, \tau)$ being specially marked.

2.2.4 An alternative formula for piecewise-constant subpopulation sizes

For demographic scenarios with piecewise-constant subpopulation sizes, we present an alternative formula for computing the truncated SFS within a constant piece. This formula has the same sample computational complexity as that described in the previous section.

Let $\mathcal{K}_t$ denote the *coalescent with killing*, a stochastic process that is closely related to the Chinese restaurant process, Hoppe's urn, and Ewens' sampling formula (Aldous, 1985; Hoppe, 1984). In particular, the coalescent with killing $\{\mathcal{K}_t\}_{t\geq 0}$ is a stochastic process whose value at time t is a *marked* partition of $\{1, \ldots, n\}$, where each partition block is marked as "killed" or "unkilled". We obtain the partition for $\mathcal{K}_t$ by dropping mutations onto the coalescent tree as a Poisson point process with rate $\frac{\theta}{2}$, and then defining an equivalence relation on $\{1, \ldots, n\}$, where $i \sim j$ if and only if i, j have coalesced by time t and there are no mutations on the branches between i and j (i.e., i and j are identical by descent). We furthermore mark the equivalence classes (i.e. partition blocks) of $\mathcal{K}_t$ that are descended from a mutation in $[0, t)$ as "killed". See Figure 2.2 for an illustration. The process $\mathcal{K}_\tau$ can also be obtained by running Hoppe's urn, or equivalently the Chinese restaurant process, forward in time (Durrett, 2008, Theorem 1.9).

Let $A_t^{\mathcal{K}}$ be the number of unkilled blocks in $\mathcal{K}_t$, so that $A_t^{\mathcal{K}}$ is a pure death process with transition rate $\lambda^{\mathcal{K}}_{i,i-1}(t) = \binom{i}{2}\alpha(t) + \frac{i\theta}{2}$ (the rate of coalescence is the number of unkilled pairs $\binom{i}{2}\alpha(t)$, and the rate of killing due to mutation is $\frac{i\theta}{2}$). Our next proposition gives a formula for the truncated conditional sample frequency spectrum given $A_\tau^{\mathcal{K}}$, i.e., $f_n^\tau(k \mid A_\tau^{\mathcal{K}} = m)$.

Proposition 1. *Consider the constant population size history* $\frac{1}{\alpha(t)} = \frac{1}{\alpha}$ *for* $t \in [0, \tau)$*, and let* $m > 0$ *and* $0 < k \leq n - m$*. The joint probability that the number of derived mutants is* k *and the number of unkilled ancestral lineages is* m*, when truncating at time* τ*, is given by*

$$\mathbb{P}_n(|\mathcal{X}^\tau| = k, A_\tau^{\mathcal{K}} = m) = \frac{\theta}{2} f_n^\tau(k \mid A_\tau^{\mathcal{K}} = m)\mathbb{P}(A_\tau^{\mathcal{C}} = m) + o(\theta),$$

where

$$f_n^\tau(k \mid A_\tau^{\mathcal{K}} = m) = \frac{2}{\alpha k} \frac{\binom{n-m}{k}}{\binom{n-1}{k}}. \tag{2.11}$$

We prove Proposition 1 in Section 2.5.5. Note that this equation does not hold for the case $k = n, m = 0$, but fortunately we do not need to consider that case in what follows below.

We can use Proposition 1 to stably and efficiently compute the terms $f_\nu^\tau(k)$, for $k \leq \nu \leq n$, as follows. We first compute the case $k < \nu = n$. Note that $\mathbb{P}_n(|\mathcal{X}^\tau| = K) = \sum_m \mathbb{P}_n(|\mathcal{X}^\tau| = K, A_\tau^{\mathcal{K}} = m)$. So for $k < n$, by Proposition 1

$$\begin{aligned} f_n^\tau(k) &= \sum_{m=1}^{n} f_n^\tau(k \mid A_\tau^{\mathcal{K}} = m) \mathbb{P}_n(A_\tau^{\mathcal{C}} = m) \\ &= \sum_{m=1}^{n} \frac{2}{\alpha k} \frac{\binom{n-m}{k}}{\binom{n-1}{k}} \mathbb{P}_n(A_\tau^{\mathcal{C}} = m). \end{aligned} \tag{2.12}$$

The sum in (2.12) contains $O(n)$ terms, so it costs $O(n^2)$ to compute $f_n^\tau(k)$ for all $k < n$. After this, we use Lemma 1 to compute $f_n^\tau(n)$, and then use Lemma 2 to compute $f_\nu^\tau(k)$ for all $1 \leq k \leq \nu < n$. Since there are $O(n^2)$ such terms, this also takes $O(n^2)$ time.

2.3 The joint SFS for multiple populations

In this section we discuss an algorithm for computing the multi-population SFS (Wakeley and Hey, 1997; Chen, 2012, 2013). We describe the algorithm in Section 2.3.1, and note how the results from Section 2.2 improve the time complexity of this algorithm. In Section 2.3.2, we focus on the special case of tree-shaped demographies, and introduce a further algorithmic speedup by replacing the coalescent with a Moran model.

Let V be the number of subpopulations in the demographic history, n the total sample size, and L the number of SFS entries to compute. Then the results from Section 2.2 improve the computational complexity of the SFS from $O(n^5V + WL)$ to $O(n^2V + WL)$, where W is a term that depends on the structure of the demographic history. In the special case of tree-shaped demographies, the algorithm from Chen (2012) gives $W = O(n^4V)$. The Moran-based speedup from Section 2.3.2 improves this to $W = O((n^2 + n\mathcal{T})V)$, with $\mathcal{T}$ being the number of matrix-vector multiplications to compute a certain matrix exponential, as described in Section 2.3.2. The complexity can be further improved to $W = O((n \log(n) + n\mathcal{T})V)$ by using the FFT to compute a convolution, but this speedup is numerically unstable; see Section 2.3.2.

The Moran-based speedup can be generalized to non-tree demographies, but the notation, implementation, and analysis of computational complexity becomes substantially more complicated. We thus leave its generalization to Chapter 3, and only describe the Moran model for tree demographies here.

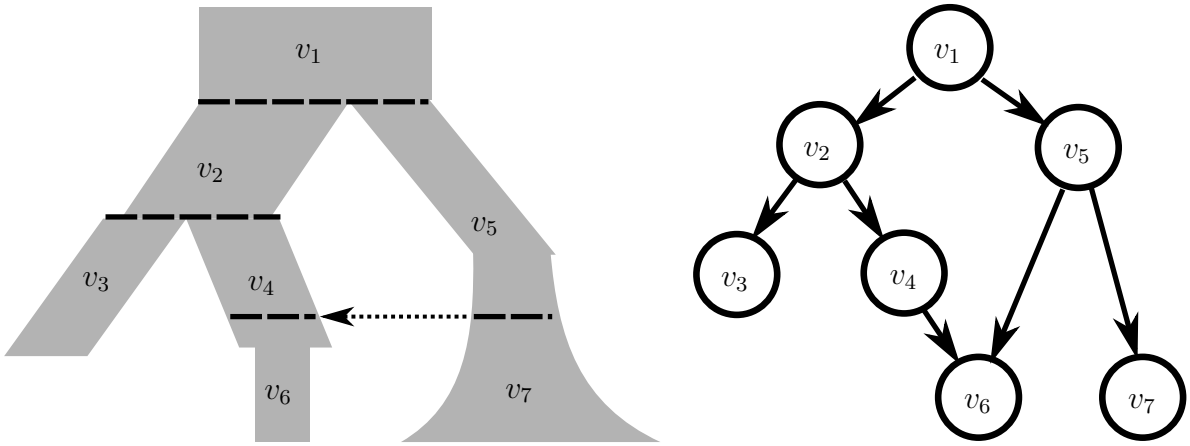


Figure 2.3: A demographic history with a pulse migration event (left), and the corresponding directed graph over the populations $v_1, \ldots, v_7$.

2.3.1 A coalescent-based dynamic program

Suppose at the present we have $\mathcal{D}$ populations, and in the ith population we observe n_i alleles. For a single point mutation, let $\mathbf{x} = (x_1, \ldots, x_{\mathcal{D}})$ denote the number of alleles that are derived in each population. We wish to compute $\xi_{\mathbf{x}}$, where $\frac{\theta}{2}\xi_{\mathbf{x}}$ is the expected number of point mutations with derived counts $\mathbf{x}$.

For demographic histories consisting of population size changes, population splits, population mergers, and pulse admixture events, Chen (2012) gave an algorithm to compute $\xi_{\mathbf{x}}$ using the truncated SFS $f_n^{\tau}(k)$ that we defined in Section 2.2.

We describe this algorithm to compute $\xi_{\mathbf{x}}$. We start by representing the population history as a directed acyclic graph (DAG), where each vertex v represents a subpopulation (Figure 2.3). We draw a directed edge from v to v' if there is gene flow from the bottom-most part of v to the top-most part of v', where "down" is the present and "up" is the ancient past. Thus, the leaf vertices correspond to the subpopulations at the present. For a vertex v in the population history graph, let $\tau_v \in (0, \infty)$ denote the length of time the corresponding population persists, and let $\alpha_v : [0, \tau_v) \to \mathbb{R}^+$ denote the inverse population size history of v. So going backwards in time from the present, $\alpha_v(t)$ gives the instantaneous rate at which two particular lineages in v coalesce, after v has existed for time t. We use $f_n^v(k)$ to denote the truncated SFS for the coalescent embedded in v, i.e., $f_n^v(k) = f_n^{\tau_v}(k)$ for a coalescent with coalescence rate $\alpha_v(t)$. Then we have

$$\xi_{\mathbf{x}} \;=\; \sum_v \sum_{m_0^v, k_0^v} f_{m_0^v}^v(k_0^v)\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v)\mathbb{P}(m_0^v) \tag{2.13}$$

where m_0^v denotes the number of lineages at the bottom of v that are ancestral to the initial sample, and k_0^v denotes the number of these lineages with a derived allele.

In order to use (2.13), we must compute $f_{m_0^v}^v(k_0^v)$ for every population v, and every value of m_0^v and k_0^v. If n is the total sample size and V the total number of vertices, then this

takes $O(n^5V)$ time using the formulas of Chen (2012). Our results from Section 2.2 improve this to $O(n^2V)$.

To use (2.13), we must also compute the terms $\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v)\mathbb{P}(m_0^v)$, for which Chen (2012) constructs a dynamic program, starting at the leaf vertices and moving up the graph. This dynamic program essentially consists of setting up a Bayesian graphical model with random variables m_0^v, k_0^v and performing belief propagation, which can be done via the sum-product algorithm ("tree-peeling") if the population graph is a tree (Pearl, 1982; Felsenstein, 1981), or via a junction tree algorithm if not (Lauritzen and Spiegelhalter, 1988).

The time complexity of the algorithm thus depends on the topological structure of the population graph. In the special case where the demographic history is a binary tree, the tree-peeling algorithm computes the values $\mathbb{P}(\mathbf{x} \mid k_0^v, m_0^v)\mathbb{P}(m_0^v)$ in $O(n^4V)$ time, since the vertex v has $O(n^2)$ possible states (k_0^v, m_0^v), so summing over the transitions between every pair of states costs $O(n^4)$. Note that Chen (2012) mistakenly states that the computation takes $O(n^3V)$ time.

To summarize, let W be the time it takes to compute (2.13) after the terms $f_m^v(k)$ have been precomputed, and let L be the number of distinct entries $\mathbf{x}$ for which we wish to compute $\xi_{\mathbf{x}}$. Then our results from Section 2.2 improve the computational complexity from $O(n^5V + WL)$ to $O(n^2V + WL)$. In the case of a binary tree the original algorithm of Chen (2012) gives $W = O(n^4V)$. In the following section, we further improve the runtime to $W = O((n^2 + n\mathcal{T})V)$ (where $\mathcal{T}$ will be the number of terms used when approximating a certain matrix exponential).

2.3.2 A Moran-based dynamic program

Here, we describe a dynamic program that improves the computational complexity of computing $\xi_{\mathbf{x}}$ for tree-shaped demographies. The main idea is to replace the backwards-in-time coalescent with a forwards-in-time Moran model.

Algorithm description

We assume the $\mathcal{D}$ populations at the present are related by a binary rooted tree with $\mathcal{D}$ leaves, where each leaf represents a population at the present, and at each internal vertex, a parent population splits into two child populations. (Note that a non-binary tree can be represented as a binary tree, with additional vertices of height 0).

Instead of working with the multi-population coalescent directly, we will consider a multi-population Moran model, in which the coalescent is embedded (Moran, 1958). In particular, let $\mathfrak{L}(v)$ denote the leaf populations descended from the population v, and let $n_v = \sum_{i \in \mathfrak{L}(v)} n_i$ be the number of present-day alleles with ancestry in v. For each population v (except the root), we construct a Moran model going *forward* in time, i.e. starting at τ_v and ending at 0. The Moran model consists of n_v lineages, each with either an ancestral or derived allele. Going forward in time, every lineage copies itself onto every other lineage at rate $\frac{1}{2}\alpha_v(t)$. Thus, the total rate of copying events is $\binom{n_v}{2}\alpha_v(t)$. Let μ_t^v denote the number of

derived alleles at time t in population v. Then the transition rate of μ_t^v when $\mu_t^v = x$ is $\lambda_{x\to x+1}(t) = \lambda_{x\to x-1}(t) = \frac{x(n_v-x)}{2}\alpha_v(t)$, since there are $x(n_v - x)$ pairs of lineages with different alleles.

The coalescent is embedded within the Moran model, because if we trace the ancestry of genetic material backwards in time in the Moran model, we obtain a genealogy with the same distribution as under the coalescent (Durrett, 2008, Theorem 1.30). Thus, we can obtain the expected number of mutations with derived counts $\mathbf{x}$, by summing over which population v the mutation occurred in:

$$\xi_{\mathbf{x}} \;=\; \sum_v \sum_{k=1}^{n_v} f_{n_v}^v(k)\mathbb{P}(\mathbf{x} \mid \mu_0^v = k). \tag{2.14}$$

Let $\mathbf{x}_v = \{x_i : i \in \mathfrak{L}(v)\}$ denote the subsample of derived allele counts in the populations descended from v. Similarly, let $\mathbf{x}_v^c = \{x_i : i \notin \mathfrak{L}(v)\}$. Then for $k \geq 1$,

$$\mathbb{P}(\mathbf{x} \mid \mu_0^v = k) \;=\; \begin{cases} \mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k), & \text{if } \mathbf{x}_v^c = \mathbf{0}, \\ 0, & \text{if } \mathbf{x}_v^c \neq \mathbf{0}. \end{cases} \tag{2.15}$$

So it suffices to compute $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k)$ for all v and k. If v is the ith leaf population, then $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k) = \mathbb{I}_{k=x_i}$. On the other hand, if v is an interior vertex with children v_1 and v_2, then

$$\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k) = \sum_{k_1=0}^{n_{v_1}} \frac{\binom{n_{v_1}}{k_1}\binom{n_{v_2}}{k-k_1}}{\binom{n_v}{k}} \mathbb{P}(\mathbf{x}_{v_1} \mid \mu_{\tau_{v_1}}^{v_1} = k_1)\mathbb{P}(\mathbf{x}_{v_2} \mid \mu_{\tau_{v_2}}^{v_2} = k - k_1), \tag{2.16}$$

where $\mathbb{P}(\mathbf{x}_{v_i} \mid \mu_{\tau_{v_i}}^{v_i})$ can be computed from

$$\mathbb{P}(\mathbf{x}_v \mid \mu_{\tau_v}^v = k) \;=\; \sum_{j=0}^{n_v} \mathbb{P}(\mathbf{x}_v \mid \mu_0^v = j)\mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k). \tag{2.17}$$

To compute the transition probability $\mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k)$, note that the transition rate matrix of μ_t^v can be written as $Q^{(v)}\alpha(t)$, where $Q^{(v)} = (q_{ij}^{(v)})_{0\leq i,j\leq n_v}$ is a $(n+1)\times(n+1)$ matrix with

$$q_{ij}^{(v)} \;=\; \begin{cases} -i(n_v - i), & \text{if } i = j, \\ \frac{1}{2}i(n_v - i), & \text{if } |j-i| = 1, \\ 0, & \text{else}, \end{cases}$$

so then the transition probability is given by the matrix exponential

$$\mathbb{P}(\mu_0^v = j \mid \mu_{\tau_v}^v = k) \;=\; (e^{Q^{(v)}\int_0^{\tau_v}\alpha_v(t)dt})_{k,j}. \tag{2.18}$$

Thus, the joint SFS $\xi_{\mathbf{x}}$ can be computed using (2.14) and (2.15), with $\mathbb{P}(\mathbf{x}_v \mid \mu_0^v = k)$ given by recursively computing (2.16), (2.17), and (2.18), in a depth-first search on the population tree (i.e., Felsenstein's tree-peeling algorithm, or the sum-product algorithm for belief propagation).

Computational complexity of Moran approach

We now consider the computational complexity associated with (2.16) and (2.17) for each vertex v. Letting $\ell_t^v(k) = \mathbb{P}(\mathbf{x}_v \mid \mu_t^v = k)$, (2.17) turns into

$$\ell_{\tau_v}^v \quad = \quad e^{\left(Q^{(v)} \int_0^{\tau_v} \alpha_v(t)dt\right)} \ell_0^v, \tag{2.19}$$

which can be efficiently computed using the method of Al-Mohy and Higham (2011). In particular, letting $A = \left(Q^{(v)} \int_0^{\tau_v} \alpha_v(t)dt\right)$ and integers $m, s \geq 1$,

$$\ell_{\tau_v}^v = e^A \ell_0^v = \left(e^{s^{-1}A}\right)^s \ell_0^v \approx \Big[\sum_{i=0}^{m-1} \frac{1}{i!}(s^{-1}A)^i\Big]^s \ell_0^v, \tag{2.20}$$

with the approximation following from truncating the Taylor series of $e^{s^{-1}A}$. Setting

$$B_j = \Big[\sum_{i=0}^{m-1} \frac{1}{i!}(s^{-1}A)^i\Big]^j \ell_0^v = \sum_{i=0}^{m-1} \frac{1}{i!}(s^{-1}A)^i B_{j-1} = \sum_{i=0}^{m-1} \frac{1}{i!} s^{-i} A(A^{i-1}B_{j-1}),$$

we have that (2.20) is equal to B_s, and B_s is evaluated in $\mathcal{T} = ms$ matrix-vector multiplications, each of which costs $O(n_v)$ by the sparsity of A. Thus, computing (2.19) costs $O(n_v\mathcal{T})$. Both m, s (and thus $\mathcal{T}$) are automatically chosen to bound the error of (2.20).

We note a similar sparse matrix exponential was used by Bryant et al. (2012), but in their context costs $O(n_v^2\mathcal{T})$, since they use the coalescent instead of the Moran model.

Next, we consider (2.16). This sum has $O(n_v)$ terms, and must be solved for $O(n_v)$ values of k, and thus costs $O(n_v^2)$ in total. We note this can be further improved to $O(n_v \log(n_v))$ by using the FFT, as in Bryant et al. (2012). In particular, letting $\tilde{\ell}_t^v(k) = \binom{n_v}{k} \ell_t^v(k)$, (2.16) can be written as a convolution

$$\tilde{\ell}_0^v = \tilde{\ell}_{\tau_{v_1}}^{v_1} * \tilde{\ell}_{\tau_{v_2}}^{v_2}, \tag{2.21}$$

which can be computed in $O(n_v \log(n_v))$ time via the fast Fourier transform (Cooley and Tukey, 1965). However, taking the Fourier transform introduces cancellation errors, due to multiplying and adding terms like e^{-ix}, and we found that converting from $\tilde{\ell}_0^v$ back to ℓ_0^v can cause these errors to blow up, due to the combinatorial factors. We thus prefer to use the naive $O(n_v^2)$ approach to compute the convolution.

The computational complexity associated with a single vertex v is thus $O(n_v^2 + n_v\mathcal{T})$. Therefore, computing the joint SFS entry $\xi_{\mathbf{x}}$ for L distinct values of $\mathbf{x}$ takes $O((n^2 + n\mathcal{T})VL)$ time for a binary population tree with arbitrary population size functions and no migration.

2.4 Runtime and accuracy results

We implemented our formulas and algorithm in Python, using the Python packages *numpy* and *scipy*. We also implemented the formulas from Chen (2012, 2013), and compared the performance of the two algorithms on simulated data.

We simulated datasets with $n \in \{2, 4, 8, \ldots, 256\}$ lineages and $\mathcal{D} \in \{2, 4, 8, \ldots, n\}$ populations at present, each containing $\frac{n}{\mathcal{D}}$ lineages. For each value of $n, \mathcal{D}$, we used the program *scrm* (Staab et al., 2015) to generate 20 random datasets, each with a demographic history that is a random binary tree.

In Figures 2.4 and 2.5, we compare the running time of the original algorithm of Chen (2012, 2013) against our new algorithm that utilizes the formulas for $f_n^\tau(k)$ presented in Section 2.2 and our new Moran-based approach described in Section 2.3.2. We find our algorithm to be orders of magnitude faster; the difference is especially pronounced as the number n of lineages grows. Note that, due to the increased running time of Chen's algorithm, we did not finish running his method for $n = 256$ and $\mathcal{D} \geq 32$.

In Figure 2.6, we compare the accuracy of the two algorithms. The figure compares the SFS entries returned by the two methods across a subset of the simulations depicted in Figure 2.4. The line $y = x$ is also plotted; points falling on the line depict the SFS entries where both methods agreed. All negative return values represent numerical errors. For $n \leq 64$ the two methods generally agree, but for larger n Chen's algorithm displays considerable numerical instability, returning extremely large positive and negative numbers.

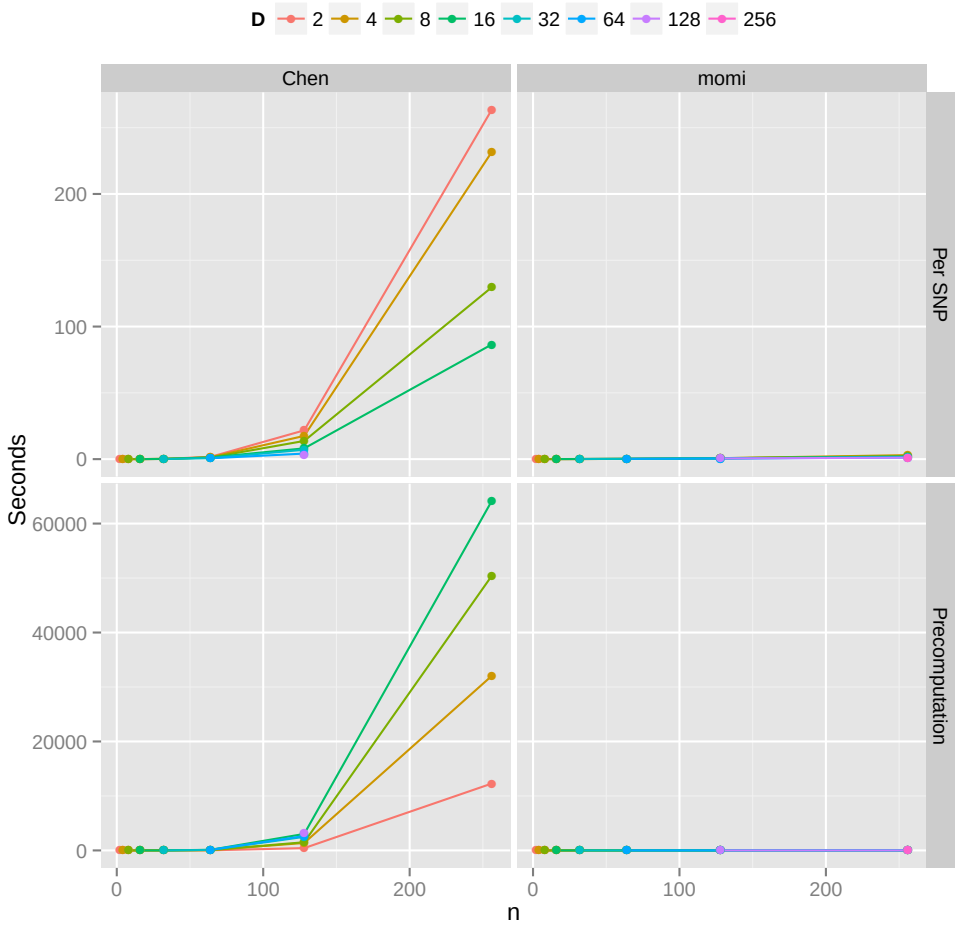


Figure 2.4: Average computation time of the joint SFS. For each combination of the sample size n and the number $\mathcal{D}$ of populations (with $\frac{n}{\mathcal{D}}$ samples per population), we generated 20 random datasets, each under a demographic history that is a random binary tree. The expected joint SFS for the resulting segregating sites were then computed using our method (*momi*) and that of Chen (2012). In the top row, we plot the average runtime per joint SFS entry, and in the bottom row, the average amount of time needed to precompute the truncated SFS for every subpopulation within each demographic history. Note the y-axis is on a different scale for each row.

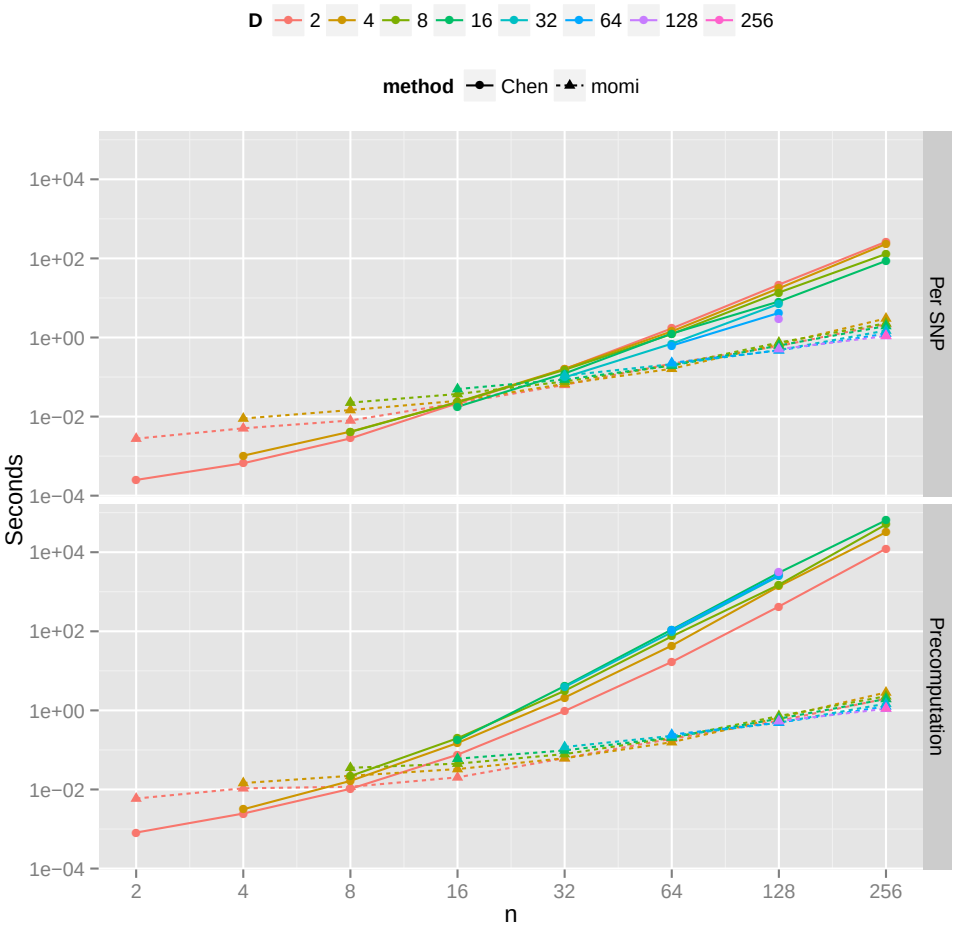


Figure 2.5: This plot is similar to Figure 2.4, but with the axes on a log-log scale, so that shorter runtimes are visible.

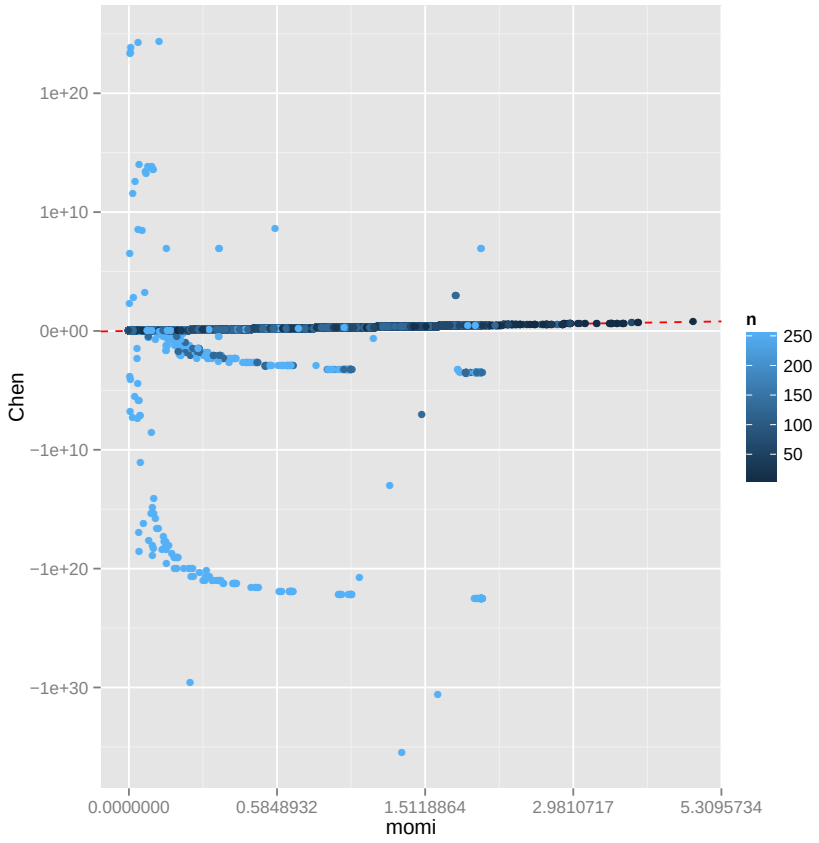


Figure 2.6: Numerical stability of the two algorithms. The plot compares the numerical values returned by our method (*momi*) and Chen's method, for the simulations described in Figure 2.4. The dashed red line represents the identity $y = x$. Note the axes were stretched by the map $z \mapsto \text{sign}(z)\log(1 + |z|)$ to adequately illustrate the observed range of numerical values. The two methods agree for $n \leq 64$, but Chen's method displays considerable numerical instability for larger n.

2.5 Proofs

In this section, we provide proofs of the mathematical results presented in earlier sections.

2.5.1 A recursion for efficiently computing $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$

We describe how to compute $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$, for all values of $m \leq \nu \leq n$, in $O(n^2)$ time. First, note that

$$\begin{aligned}
&\mathbb{P}_{\nu-1}(A_\tau^{\mathcal{C}} = m) \\
&= \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1, \{\nu\} \in \mathcal{C}_\tau) + \mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m, \{\nu\} \notin \mathcal{C}_\tau) \\
&= \frac{(m+1)p_{\nu,m+1}^{1,1}}{\binom{\nu}{1}}\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1) + \left(1 - \frac{mp_{\nu,m}^{1,1}}{\binom{\nu}{1}}\right)\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m) \\
&= \frac{(m+1)(m)}{\nu(\nu-1)}\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1) + \left(1 - \frac{m(m-1)}{\nu(\nu-1)}\right)\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m).
\end{aligned}$$

Rearranging, we get the recursion

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m) = \frac{1}{1 - \frac{m(m-1)}{\nu(\nu-1)}}\left[\mathbb{P}_{\nu-1}(A_\tau^{\mathcal{C}} = m) - \frac{(m+1)(m)}{\nu(\nu-1)}\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m+1)\right] \tag{2.22}$$

with base cases

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2}\int_0^\tau \alpha(t)dt}.$$

So after solving $\int_0^\tau \alpha(t)dt$, we can use the recursion and memoization to solve for all of the $O(n^2)$ terms $\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = m)$ in $O(n^2)$ time. In particular, in the case of constant population size, $\alpha(t) = \alpha$, the base case is given by

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2}\alpha\tau},$$

and in the case of an exponentially growing population size, $\alpha(t) = \alpha(\tau)e^{\beta(\tau-t)}$, the base case is given by

$$\mathbb{P}_\nu(A_\tau^{\mathcal{C}} = \nu) = e^{-\binom{\nu}{2}\alpha(\tau)(e^{\beta\tau} - \frac{1}{\beta})}.$$

2.5.2 Proof of Lemma 1

Let T_{MRCA} denote the time to the most recent common ancestor of the sample. We first note that

$$f_n^\tau(n) = \tau - \mathbb{E}_n[T_{\text{MRCA}} \wedge \tau],$$

since the branch length subtending the whole sample is the time between τ and T_{MRCA}.

Next, note that $\frac{\theta}{2}\mathbb{E}_n[T_{\text{MRCA}} \wedge \tau]$ is equal to the number of polymorphic mutations in $[0, \tau)$ where the individual "1" is derived. This is because, as we trace the ancestry of "1" backwards in time, all mutations hitting the lineage below T_{MRCA} are polymorphic, while all mutations hitting above T_{MRCA} are monomorphic.

The expected number of polymorphisms with "1" derived also equals $\frac{\theta}{2}\sum_{k=1}^{n-1}\frac{k}{n}f_n^\tau(k)$, since if a mutation has k derived leaves, the chance that "1" is in the derived set is $\frac{k}{n}$. Thus,

$$\mathbb{E}_n[T_{\text{MRCA}} \wedge \tau] \;=\; \sum_{k=1}^{n-1}\frac{k}{n}f_n^\tau(k),$$

which completes the proof.

2.5.3 Proof of Lemma 2

We first note that

$$\begin{aligned}&\mathbb{P}_n(\mathcal{X}^\tau = \{1,\ldots,k\})\\ &\quad= \mathbb{P}_{n+1}(\mathcal{X}^\tau = \{1,\ldots,k\}) + \mathbb{P}_{n+1}(\mathcal{X}^\tau = \{1,\ldots,k,n+1\}).\end{aligned}$$

By exchangeability, we have $\mathbb{P}_n(\mathcal{X}^\tau = K) = \frac{\theta}{2}\frac{f_n^\tau(|K|)}{\binom{n}{|K|}} + o(\theta)$ for all $K \subseteq \{1,\ldots,n\}$, so

$$\frac{1}{\binom{n}{k}}f_n^\tau(k) = \frac{1}{\binom{n+1}{k}}f_{n+1}^\tau(k) + \frac{1}{\binom{n+1}{k+1}}f_{n+1}^\tau(k+1).$$

Multiplying both sides by $\binom{n}{k}$ gives

$$f_n^\tau(k) \;=\; \frac{n-k+1}{n+1}f_{n+1}^\tau(k) + \frac{k+1}{n+1}f_{n+1}^\tau(k+1).$$

2.5.4 Proof of Lemma 3

Let $\alpha^*(t)$ denote the inverse population size history given by

$$\alpha^*(t) \;=\; \begin{cases}\alpha(t) & \text{if } t < \tau\\ \infty & \text{if } t \geq \tau.\end{cases}$$

So the demographic history with population size $\frac{1}{\alpha^*(t)}$ agrees with the original history up to time τ, at which point the population size drops to 0, and all lineages instantly coalesce into a single lineage with probability 1.

Let $T_{m,*}$ denote the amount of time there are m ancestral lineages for the coalescent with size history $\frac{1}{\alpha^*(t)}$. Similarly, let ξ_k^* denote the SFS under the size history $\frac{1}{\alpha^*(t)}$. Then from the result of Polanski and Kimmel (2003),

$$\xi_k^* = \sum_{m=2}^{n} W_{n,k,m} \mathbb{E}_m[T_{m,*}].$$

Note that for $m > 1$, we almost surely have $T_{m,*} = T_{m,*}^{\tau}$, i.e. the intercoalescence time equals its truncated version, since all lineages coalesce instantly at τ with probability 1. Thus, $\mathbb{E}_m[T_{m,*}] = \mathbb{E}_m[T_{m,*}^{\tau}]$. Similarly, for $k < n$, $\xi_k^* = f_n^{*\tau}(k)$, i.e. the SFS equals the truncated SFS, because the probability of a polymorphic mutation occurring in $[\tau, \infty)$ is 0.

Finally, note that $\mathbb{E}_m[T_{m,*}^{\tau}] = \mathbb{E}_m[T_m^{\tau}]$ and $f_n^{*\tau}(k) = f_n^{\tau}(k)$, because $\alpha(t)$ and $\alpha^*(t)$ are identical on $[0, \tau)$.

2.5.5 Proof of Proposition 1

We start by showing that $\mathbb{P}_n(A_\tau^{\mathcal{K}} = m) = \mathbb{P}_n(A_\tau^{\mathcal{C}} = m) + O(\theta)$. Let $T_i(\mathcal{K}) = \int_0^\tau \mathbb{I}_{A_t^{\mathcal{K}}=i} dt$ denote the amount of time where $\mathcal{K}$ has i unkilled lineages. Let p denote the probability density function. For $(t_n, \ldots, t_m)$ with $\sum t_i = \tau$, we have

$$\begin{aligned}
& p(T_n(\mathcal{K}) = t_n, \ldots, T_m(\mathcal{K}) = t_m) \\
= \; & e^{-\lambda_{m,m-1}^{\mathcal{K}} t_m} \prod_{i=m+1}^{n} \lambda_{i,i-1}^{\mathcal{K}} e^{-\lambda_{i,i-1}^{\mathcal{K}} t_i} \\
= \; & e^{-\left(\binom{m}{2}\alpha + \frac{m\theta}{2}\right) t_m} \prod_{i=m+1}^{n} \left(\binom{i}{2}\alpha + \frac{i\theta}{2} \right) e^{-\left(\binom{i}{2}\alpha + \frac{i\theta}{2}\right) t_i} \\
= \; & e^{-\binom{m}{2}\alpha t_m} \prod_{i=m+1}^{n} \binom{i}{2} \alpha e^{-\binom{i}{2}\alpha t_i} + O(\theta) \\
= \; & p(T_n = t_n, \ldots, T_m = t_m) + O(\theta),
\end{aligned}$$

and so

$$\begin{aligned}
\lim_{\theta \to 0} \mathbb{P}_n(A_\tau^{\mathcal{K}} = m) &= \lim_{\theta \to 0} \int_{\sum t_i = \tau} p(T_n(\mathcal{K}) = t_n, \ldots, T_m(\mathcal{K}) = t_m) d\mathbf{t} \\
&= \int_{\sum t_i = \tau} p(T_n = t_n, \ldots, T_m = t_m) d\mathbf{t} \\
&= \mathbb{P}_n(A_\tau^{\mathcal{C}} = m).
\end{aligned}$$

where we can exchange the limit and the integral by the Bounded Convergence Theorem, because $p(T_n(\mathcal{K}) = t_n, \ldots, T_m(\mathcal{K}) = t_m) \leq \prod_{i=m+1}^{n} \left(\binom{i}{2}\alpha + \frac{i}{2} \right)$ for $\theta \leq 1$.

Thus we have

$$\begin{aligned}\mathbb{P}_n(|\mathcal{X}^\tau| = k, A^{\mathcal{K}}_\tau = m) &= \mathbb{P}_n(|\mathcal{X}^\tau| = k \mid A^{\mathcal{K}}_\tau = m)\mathbb{P}_n(A^{\mathcal{K}}_\tau = m) \\ &= \left(\frac{\theta}{2} f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m) + o(\theta)\right)\left(\mathbb{P}_n(A^{\mathcal{C}}_\tau = m) + O(\theta)\right) \\ &= \frac{\theta}{2} f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m)\mathbb{P}_n(A^{\mathcal{C}}_\tau = m) + o(\theta),\end{aligned}$$

which proves the first part of the proposition.

We next solve for $f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m)$, the first order Taylor series coefficient for $\mathbb{P}_n(|\mathcal{X}^\tau| = k \mid A^{\mathcal{K}}_\tau = m)$ in the mutation rate $\frac{\theta}{2}$.

When there are i unkilled lineages, the probability that the next event is a killing event is $\frac{\theta}{\alpha(i-1)+\theta} = \frac{\theta}{\alpha(i-1)} + o(\theta)$. Given that the event is a killing, the chance that the killed lineage has k leaf descendants is $p^{k,1}_{n,i}$. So summing over i, and dividing out the mutation rate $\frac{\theta}{2}$, we get

$$\begin{aligned}f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m) &= \frac{2}{\alpha}\sum_{i=m+1}^{n-k+1} \frac{1}{i-1} p^{k,1}_{n,i} \\ &= \frac{2}{\alpha}\sum_{i=m+1}^{n-k+1} \frac{1}{i-1}\frac{\binom{n-k-1}{i-2}}{\binom{n-1}{i-1}} \\ &= \frac{2}{\alpha}\sum_{i=m+1}^{n-k+1} \frac{1}{i-1}\frac{(n-k-1)!(i-1)!(n-i)!}{(i-2)!(n-k-i+1)!(n-1)!} \\ &= \frac{2(n-k-1)!}{\alpha(n-1)!}\sum_{i=m+1}^{n-k+1} \frac{(n-i)!}{(n-k-i+1)!} \\ &= \frac{2(n-k-1)!}{\alpha(n-1)!}\sum_{j=0}^{n-k-m} \frac{(j+k-1)!}{j!} \\ &= \frac{2}{\alpha k\binom{n-1}{k}}\sum_{j=0}^{n-k-m}\binom{j+k-1}{j} \\ &= \frac{2}{\alpha k}\frac{\binom{n-m}{k}}{\binom{n-1}{k}},\end{aligned}$$

where we made the change of variables $j = n-k-i+1$, and where the final line follows from repeated application of the combinatorial identity $\binom{a}{b} = \binom{a-1}{b} + \binom{a-1}{b-1}$.

Alternative proof for $f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m)$ via the Chinese Restaurant Process

We sketch an alternative proof of the expression for $f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m)$, using the Chinese Restaurant Process.

Consider the coalescent with killing going forward in time (towards the present), and only looking at it when the number of individuals increases. Then when there are i lineages, a new mutation occurs with probability $\frac{\theta}{\alpha i+\theta} = \frac{\theta/\alpha}{i+\theta/\alpha}$, and each lineage branches with probability $\frac{\alpha}{\alpha i+\theta} = \frac{1}{i+\theta/\alpha}$. Thus, conditional on $A^{\mathcal{K}}_\tau = m$, the distribution on $\mathcal{K}_\tau$ is given by a Chinese Restaurant Process (Aldous, 1985), starting with m tables each with 1 person, and with new tables founded with parameter θ/α.

Let $(x)_{i\uparrow} = x(x+1)\cdots(x+i-1)$ denote the rising factorial. If there is a single mutation with k descendants, then there are $\binom{n-m}{k}$ ways to pick which of the $n-m$ events involve mutant lineages. The probability of a particular such ordering is

$$\frac{\theta}{\alpha}\frac{(1)_{k\uparrow}(m)_{n-k-m\uparrow}}{(m+\theta/\alpha)_{n-m\uparrow}} = \frac{\theta}{\alpha}\frac{(k-1)!(n-k-1)!/m!}{(n-1)!/m!} + o(\theta).$$

Summing over all $\binom{n-m}{k}$ orderings, and dividing by $\frac{\theta}{2}$, yields

$$f^\tau_n(k \mid A^{\mathcal{K}}_\tau = m) = \frac{2}{\alpha}\binom{n-m}{k}\frac{(k-1)!(n-k-1)!/m!}{(n-1)!/m!}.$$

Chapter 3

A Moran model for the SFS with admixture

In this chapter, we extend the multipopulation Moran model of Chapter 2 to include pulse migrations between populations.

As before, we construct a graph $\mathcal{G}$ where each vertex v represents an historical subpopulation. Due to admixture events, $\mathcal{G}$ is a DAG (directed acyclic graph). We embed a Moran model within $\mathcal{G}$, but due to admixture, it will be more convenient for us to use the *lookdown construction*, a variant of the Moran model with a countably infinite number of lineages (Donnelly and Kurtz, 1996; Donnelly et al., 1999). However, note that we will only keep track of a finite number of lineages at a time, as we can implicitly integrate over the infinite remaining lineages in the lookdown model.

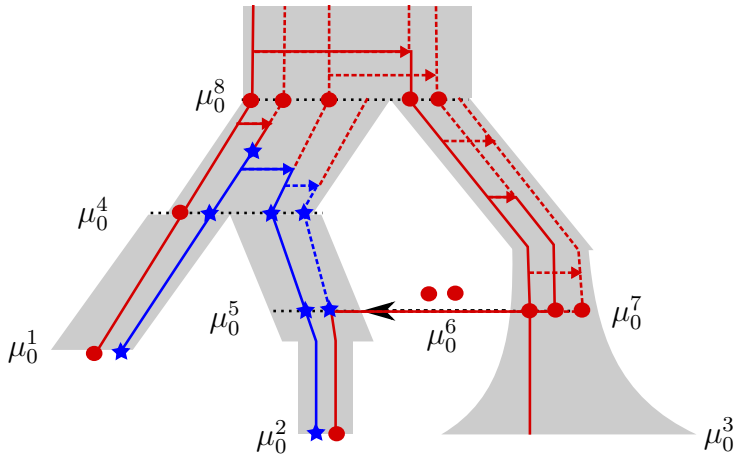
In addition to the DAG $\mathcal{G}$, we will also construct a tree $\mathcal{T}$ over the demographic events. We call $\mathcal{T}$ an "event tree". In Algorithm 1, we define a DP (dynamic program) over $\mathcal{T}$, which can be used to compute the SFS $\xi_{\mathbf{x}}$ (Theorem 1). In fact, $\mathcal{T}$ is essentially a junction tree of $\mathcal{G}$, and Algorithm 1 the junction tree algorithm from Bayesian graphical models (Lauritzen and Spiegelhalter, 1988). However, our case is distinguished from the usual junction tree algorithm, because the event tree $\mathcal{T}$ is rooted, and because we are not computing a likelihood $\mathbb{P}(\mathbf{x})$, but instead the expected count $\xi_{\mathbf{x}}$ of mutations with $\mathbf{x}$ derived alleles.

In practice, we will also need the normalizing constant $\|\xi\|_1 = \sum_{\mathbf{x}} \xi_{\mathbf{x}}$, because the probability a mutation has configuration $\mathbf{x}$ is $\frac{\xi_{\mathbf{x}}}{\|\xi\|_1}$. At first glance, computing $\|\xi\|_1$ requires computing all $O(n^{\mathcal{D}})$ SFS entries $\xi_{\mathbf{x}}$, but in Corollary 1 we show how Algorithm 1 can compute $\|\xi\|_1$ in the same time as $O(1)$ entries $\mathbf{x}$. In fact, Algorithm 1 can compute a number of interesting quantities, including $\mathbb{E}[T_{\text{MRCA}}]$, F_{ST}, and expectations of many summary statistics of the SFS.

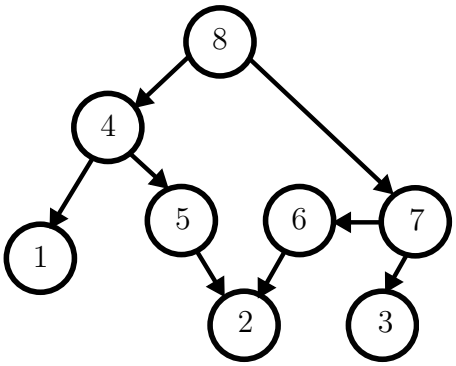
Finally, we will examine how well we can infer complicated demographic histories with the SFS. We fit demographic histories by maximizing a composite likelihood, using gradient descent and automatic differentiation. As an example, we consider a toy demography, loosely based on human history, with 6 populations and 18 parameters, including pulse migrations and exponential growth.

We delay all proofs to Section 3.5.

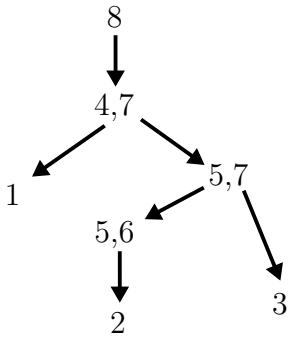
3.1 Notation and background



(a) A demographic history with lookdown Moran model. μ_0^v is the number of derived alleles (blue stars) at the bottom of population v. The observed configuration is $\mathbf{x} = (\mu_0^1, \mu_0^2, \mu_0^3) = (1, 1, 0)$. The coalescent is in solid lines.



(b) The DAG $\mathcal{G}$. Each vertex v corresponds to a population, with a path from v to w iff w has ancestry in v.



(c) The event tree $\mathcal{T}$. Each internal $\mathbf{v}$ corresponds to a join or split event, and is labeled by a subset of contemporaneous populations.

Figure 3.1: Summary of important notation and data structures.

3.1.1 The Demographic DAG

As in the previous chapter, we represent the demographic history as a directed acyclic graph $\mathcal{G}$ (Figure 3.1). The vertices of $\mathcal{G}$ are the populations of the demographic history. The leaf vertices are the sampled populations, which we label as $\{1, \ldots, \mathcal{D}\}$. For a population vertex

v, $\frac{1}{\alpha_v(t)}$ denotes the (scaled) population size at time t within v, with $t = 0$ the "bottom" (most recent time) in v and $t = \tau_v$ the "top" (most ancient time).

We can view $\mathcal{G}$ as containing two kinds of events:

1. Split events: going backwards in time, a single population w splits into two parent populations v_1 and v_2. A lineage in w passes through v_1 with probability q_1, and through v_2 with probability $q_2 = 1 - q_1$. $\mathcal{G}$ contains the directed edges $v_1 \to w$ and $v_2 \to w$.

2. Join events: going backwards in time, two populations w_1 and w_2 find a common ancestor v. $\mathcal{G}$ contains the directed edges $v \to w_1$ and $v \to w_2$.

Other demographic events may be represented as split and join events. In particular, a pulse migration is a population split immediately below a population join (for example, in Figure 3.1, the pulse migration is represented as the split $2 \to 5, 6$ and the join $6, 3 \to 7$). Likewise, k-ary join events, where $w_1, \ldots, w_k$ find a common parent v, can be viewed as a sequence of $k - 1$ pairwise join events.

3.1.2 The event tree

We now define an "event tree" $\mathcal{T}$ on top of the DAG $\mathcal{G}$, with each node $\mathbf{v} \in V(\mathcal{T})$ a demographic event *identified* with a set of populations, so that we write $\mathbf{v} = \{v_1, \ldots, v_{|\mathbf{v}|}\} \subset V(\mathcal{G})$. $\mathcal{T}$ is essentially a *junction tree* over $\mathcal{G}$ (Lauritzen and Spiegelhalter, 1988), except that $\mathcal{T}$ is rooted, whereas junction trees are usually defined to be unrooted.

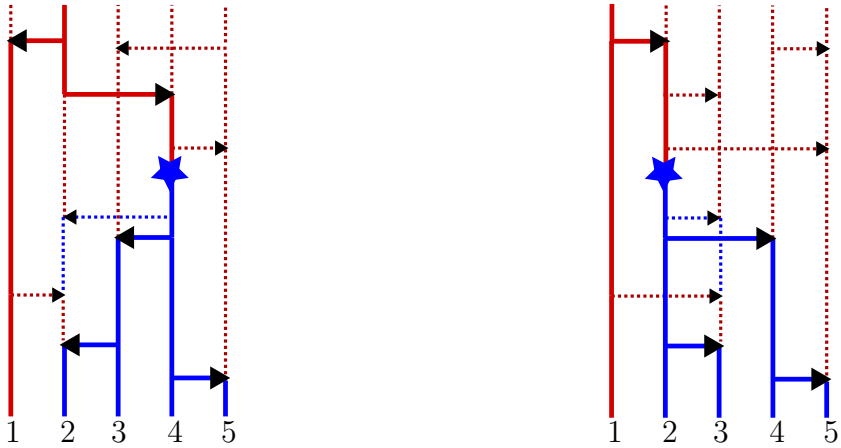
We illustrate an example of $\mathcal{T}$ in Figure 3.1. Note the junction tree of $\mathcal{G}$ is not generally unique. We choose a particular tree by constructing $\mathcal{T}$ from bottom to top, processing the events in the order of their time from the present:

1. (Leaf "event") For each $d \in \{1, \ldots, \mathcal{D}\}$, define the leaf "event" $\mathbf{v} := \{d\}$.

2. (Split event) If $\mathbf{v}$ is a split event, where a child population w splits into parents v_1, v_2, let $\mathbf{w}$ be the topmost event with $w \in \mathbf{w}$. Then we set $\mathbf{w}$ as the sole child event of $\mathbf{v}$, and set $\mathbf{v} := \mathbf{w} \cup \{v_1, v_2\} \setminus \{w\}$.

3. (Join event) If $\mathbf{v}$ is a join event, where child populations w_1, w_2 join into a parent population v, then $\mathbf{v}$ may have 1 or 2 child events. In particular, let $\mathbf{w}_1$ be the topmost event with $w_1 \in \mathbf{w}_1$, and likewise for $\mathbf{w}_2$ (with $\mathbf{w}_1 = \mathbf{w}_2$ possibly). Then we set $\mathbf{w}_1, \mathbf{w}_2$ as the children of $\mathbf{v}$, and set $\mathbf{v} := \mathbf{w}_1 \cup \mathbf{w}_2 \cup \{v\} \setminus \{w_1, w_2\}$.

We denote the child events of $\mathbf{v}$ as $C_{\mathcal{T}}(\mathbf{v})$. We denote the root of $\mathcal{T}$ by $\boldsymbol{\rho} = \{\rho\}$, with ρ the root ancestral population of $\mathcal{G}$.



(a) The standard Moran model, with copying in both directions.

(b) The lookdown Moran model, with copying in only one direction.

Figure 3.2: Standard and lookdown Moran models. The coalescent is in solid lines and identical in both 3.2a and 3.2b. In fact, 3.2b is "coupled" to 3.2a, and was obtained by swapping lineage positions above copying events going the "wrong" way.

3.1.3 Lookdown Moran model

Within $\mathcal{G}$, we embed a *lookdown construction* (Donnelly and Kurtz, 1996; Donnelly et al., 1999), a variant of the Moran model containing a *countably infinite* number of lineages, each with a unique label in $\mathbb{Z}_+$. Going forward in time, alleles copy onto each other at rate $\alpha_v(t)$ per pair of lineages. However, copying only occurs in one direction, from lineages with lower labels to lineages with higher labels. Contrast this with the standard Moran model, where copying happens in both directions at rate $\frac{\alpha_v(t)}{2}$. However, both versions of the Moran model have sample genealogies distributed as the coalescent, and thus have equivalent sampling probabilities (Figure 3.2).

We assign the $(n_1, \ldots, n_{\mathcal{D}})$ sampled lineages to the lowest labels $\{1, 2, \ldots, n_{\text{tot}}\}$, where $n_{\text{tot}} \equiv \sum_d n_d$, and assign the remaining unsampled lineages to labels $\{n_{\text{tot}} + 1, n_{\text{tot}} + 2, \ldots\}$. For population $v \in V(\mathcal{G})$ and time $t \in [0, \tau_v]$ (with $t = 0$ the bottom and $t = \tau_v$ the top of v), we denote the labeled alleles by $\mathcal{M}_{v,t} = (\mathcal{M}_{v,t,(1)}, \mathcal{M}_{v,t,(2)}, \ldots)$ with $\mathcal{M}_{v,t,(i)}$ the ith lowest label at v and its allele at time t. Note the labels at v may be random because of split (admixture) events, where each labeled allele independently chooses its ancestral population.

Now let $\mu_t^{v,m} = \sum_{i=1}^{m} \mathbb{I}_{\{\mathcal{M}_{v,t,(i)} \text{ derived}\}}$ denote the number of derived alleles among the first m lineages at v, t. Let n_v be the number of sampled alleles with potential ancestry in v. We will only need to keep track of the first n_v lineages in v, and thus set $\mu_t^v = \mu_t^{v,n_v}$.

Intuitively, we only need to keep track of the first n_v alleles because there are at most n_v labels $\leq n_{\text{tot}}$ at v, and the data are conditionally independent of the higher labels due to the lookdown property. Thus the lookdown model allows computational savings by keeping track of only the n_v lowest alleles at each v. For example, in Figure 3.1, notice how extra

non-ancestral lineages appear due to the split event at vertex 2, but are then removed in the join event at vertex 8.

For event $\mathbf{v} = (v_1, \ldots, v_k) \in V(\mathcal{T})$ and corresponding times $\mathbf{t} = (t_1, \ldots, t_k)$, we define $\mathcal{M}_{\mathbf{v},\mathbf{t}} = (\mathcal{M}_{v_1,t_1}, \ldots, \mathcal{M}_{v_k,t_k})$ as the labeled alleles at $\mathbf{v}, \mathbf{t}$, and define $\boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}} = (\mu_{t_1}^{v_1}, \ldots, \mu_{t_k}^{v_k})$ as the derived allele counts at $\mathbf{v}, \mathbf{t}$.

3.1.4 SFS and likelihoods

Let $\mathbf{x} = (x_1, \ldots, x_{\mathcal{D}}) \in \mathbb{Z}_+^{\mathcal{D}}$ be a configuration of counts in the leaf populations, and $\mathcal{B}_{\mathbf{x}}$ the total branch length in the sample genealogy with $\mathbf{x}$ descendants. The SFS $\xi_{\mathbf{x}}$ is defined by $\xi_{\mathbf{x}} = \mathbb{E}[\mathcal{B}_{\mathbf{x}}]$. Equivalently, $\xi_{\mathbf{x}}$ is the expected number of mutations with derived allele counts $\mathbf{x}$, per unit mutation rate.

Our main result is a dynamic program that returns $\xi_{\mathbf{x}}$, via computing an intermediate "partial SFS" $\xi_{\mathbf{x}}^{\mathbf{v}}$, that we now define. For event $\mathbf{v} \in V(\mathcal{T})$, let $\mathbf{x}_{\mathbf{v}} = \sum_{d \in \mathfrak{L}(\mathbf{v})} x_d \mathbf{e}_d$ be the subsample at $\mathfrak{L}(\mathbf{v})$, the leaves of $\mathbf{v}$ (where $\mathbf{e}_d$ is the unit vector with 1 at coordinate d). Now cut the demography at $\boldsymbol{\tau}_{\mathbf{v}} = \sum_{v \in \mathbf{v}} \tau_v \mathbf{e}_v$, the topmost times of $\mathbf{v}$. Keeping the connected component with $\mathbf{v}$, let $\mathcal{B}_{\mathbf{x}_{\mathbf{v}}}^{\boldsymbol{\tau}_{\mathbf{v}}}$ denote the branch length in this component with $\mathbf{x}_{\mathbf{v}}$ descendants in the sample. Then we define

$$\xi_{\mathbf{x}}^{\mathbf{v}} = \mathbb{E}[\mathcal{B}_{\mathbf{x}_{\mathbf{v}}}^{\boldsymbol{\tau}_{\mathbf{v}}}].$$

So $\xi_{\mathbf{x}}^{\mathbf{v}}$ is the expected number of mutations, occurring below $\boldsymbol{\tau}_{\mathbf{v}}$, with sampled alleles $\mathbf{x}_{\mathbf{v}}$, per unit mutation rate. Our algorithm computes $\xi_{\mathbf{x}}^{\mathbf{v}}$ for every event $\mathbf{v} \in V(\mathcal{T})$, finally yielding the desired SFS $\xi_{\mathbf{x}} = \xi_{\mathbf{x}}^{\boldsymbol{\rho}}$ at the root.

To compute $\xi_{\mathbf{x}}^{\mathbf{v}}$, we will need to consider the likelihood that the allele counts at $\mathbf{v}$ give rise to subsample $\mathbf{x}_{\mathbf{v}}$ at the leaves. We call this quantity a *partial likelihood*. More specifically, for $\mathbf{v} \in V(\mathcal{T})$ and times $\mathbf{t} = \sum_{v \in \mathbf{v}} t_v \mathbf{e}_v$, we define the partial likelihood as

$$\ell_{\boldsymbol{\mu},\mathbf{x}}^{\mathbf{v},\mathbf{t}} = \mathbb{P}_{\theta=0}(\mathbf{x}_{\mathbf{v}} \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}} = \boldsymbol{\mu})$$

the probability of $\mathbf{x}_{\mathbf{v}}$ given $\boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}} = \boldsymbol{\mu}$ and no additional mutation below $\mathbf{t}$.

3.2 Theoretical Results

We now present a dynamic program (Algorithm 1) that can compute the SFS $\xi_{\mathbf{x}}$ (Theorem 1). In fact, Algorithm 1 efficiently computes many linear statistics of the SFS (Corollary 1), such as the total branch length $\sum_{\mathbf{x}} \xi_{\mathbf{x}}$ and $\mathbb{E}[T_{\mathrm{MRCA}}]$.

3.2.1 Dynamic program for the SFS

In this subsection, we drop the dependence on $\mathbf{x}$, so that

$$\ell_{\boldsymbol{\mu}}^{\mathbf{v},\mathbf{t}} = \ell_{\boldsymbol{\mu},\mathbf{x}}^{\mathbf{v},\mathbf{t}}$$
$$\xi^{\mathbf{v}} = \xi_{\mathbf{x}}^{\mathbf{v}}.$$

Algorithm 1 Dynamic program for SFS

procedure DP($\ell^1, \ldots, \ell^{\mathcal{D}}$) ▷ $\ell^d = [\mathbb{P}(x_d \mid \mu_0^d = i)]_{0 \le i \le n_d} \in \mathbb{R}^{n_d+1}$
  for events $\mathbf{v}$ in DepthFirstSearch($V(\mathcal{T})$) **do**
    if $\mathbf{v} = \{d\}$ is leaf **then** ▷ $\ell^{\mathbf{v},\mathbf{0}} = [\mathbb{P}(\mathbf{x}_\mathbf{v} \mid \boldsymbol{\mu}_\mathbf{0}^\mathbf{v} = \boldsymbol{\mu})]_{\boldsymbol{\mu} \in \prod_{v \in \mathbf{v}} \{0,\ldots,n_v\}}$
      $\ell^{\mathbf{v},\mathbf{0}} \leftarrow \ell^d$
    else if $\mathbf{v}$ is split event **then**
      $\ell^{\mathbf{v},\mathbf{0}} \leftarrow$ (3.4), (3.3)
    else if $\mathbf{v}$ is join event **then**
      if $|C_\mathcal{T}(\mathbf{v})| = 1$ **then** ▷ $C_\mathcal{T}(\mathbf{v})$ the child events of $\mathbf{v}$
        $\ell^{\mathbf{v},\mathbf{0}} \leftarrow$ (3.5), (3.3)
      else if $|C_\mathcal{T}(\mathbf{v})| = 2$ **then**
        $\ell^{\mathbf{v},\mathbf{0}} \leftarrow$ (3.6), (3.3)
      end if
    end if ▷ Computed the partial likelihood $\ell^{\mathbf{v},\mathbf{0}}$
    $\xi^\mathbf{v} \leftarrow$ (3.1), (3.2) ▷ $\xi^\mathbf{v}$ the partial SFS
  end for
  return $\xi^{\boldsymbol{\rho}}$ ▷ $\boldsymbol{\rho}$ the root event
end procedure

Algorithm 1 defines a dynamic program (DP) over $\ell_{\boldsymbol{\mu}}^{\mathbf{v},\mathbf{0}}, \xi^\mathbf{v}$, using equations (3.1), (3.2), (3.3), (3.4), (3.5), (3.6) to be defined shortly. For appropriate inputs the DP computes the SFS:

Theorem 1. *For polymorphic* $\mathbf{x} = (x_1, \ldots, x_\mathcal{D}) \neq \mathbf{0}, \mathbf{n}$ *and leaf population* $d \in \{1, \ldots, \mathcal{D}\}$, *let* $\mathbf{e}_{x_d} = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^{(n_d+1)}$ *have 1 at coordinate* x_d *and 0 elsewhere. Then*

$$\xi_\mathbf{x} = \mathrm{DP}(\mathbf{e}_{x_1}, \ldots, \mathbf{e}_{x_\mathcal{D}}).$$

We now present the formulas used by Algorithm 1, in a series of lemmas that also prove Theorem 1. We start with a formula to compute $\xi^\mathbf{v}$ from $\ell_{\boldsymbol{\mu}}^{\mathbf{v},\mathbf{0}}$ and the partial SFS at the child events $C_\mathcal{T}(\mathbf{v})$.

Lemma 4. *For* $\mathbf{v} \in V(\mathcal{T})$ *and* $\mathbf{w} = \bigcup C_\mathcal{T}(\mathbf{v}) = \{w \in V(\mathcal{G}) \mid w \in \mathbf{w}', \mathbf{w}' \in C_\mathcal{T}(\mathbf{v})\}$,

$$\xi^\mathbf{v} = \xi^\mathbf{w} + \sum_{v \in \mathbf{v} \setminus \mathbf{w}} \sum_{k=1}^{n_v} f_{n_v}^v(k) \ell_{k\mathbf{e}_v}^{\mathbf{v},\mathbf{0}} \tag{3.1}$$

with $f_{n_v}^v(k)$ *being the truncated SFS within population* v *(as given by* (2.4) *and* (2.9), *Chapter 2), and with* $\xi^\mathbf{w}$ *given by*

$$\xi^\mathbf{w} = \begin{cases} \sum_{i \neq j} \xi^{\mathbf{w}_i} \prod_{d \in \mathfrak{L}(\mathbf{w}_j)} \ell_\mathbf{0}^{\{d\},\mathbf{0}}, & \text{if } C_\mathcal{T}(\mathbf{v}) = \{\mathbf{w}_1, \mathbf{w}_2\}, \\ \xi^{\mathbf{w}_1}, & \text{if } C_\mathcal{T}(\mathbf{v}) = \{\mathbf{w}_1\}, \\ 0, & \text{if } \mathbf{v} \text{ is a leaf of } \mathcal{T}. \end{cases} \tag{3.2}$$

Now we need formulas for computing the partial likelihood $\ell^{\mathbf{v},\mathbf{0}}$ at the bottom of $\mathbf{v}$. These will be given by (3.4), (3.5), (3.6), which will in turn depend on partial likelihoods $\ell^{\mathbf{w},\mathbf{t}}$ at child events $\mathbf{w} \in C_{\mathcal{T}}(\mathbf{v})$ and times $\mathbf{t} = \sum_{w \in \mathbf{w} \setminus \mathbf{v}} \tau_w \mathbf{e}_w$. The next formula allows us to obtain $\ell^{\mathbf{w},\mathbf{t}}$ from the previously computed likelihoods $\ell^{\mathbf{w},\mathbf{0}}$ at the bottom of $\mathbf{w}$:

Lemma 5. *Let $v \in \mathbf{v} \in V(\mathcal{T})$, $\boldsymbol{\mu}_{-v}$ a configuration of alleles on $\mathbf{v} \setminus \{v\}$, and $\mathbf{t}_{-v}$ a collection of times on $\mathbf{v} \setminus \{v\}$. Then for $\boldsymbol{\mu} = k\mathbf{e}_v + \boldsymbol{\mu}_{-v}$ and $\mathbf{t} = \tau_v \mathbf{e}_v + \mathbf{t}_{-v}$,*

$$\ell^{\mathbf{v},\mathbf{t}}_{\boldsymbol{\mu}} = \sum_{j=0}^{n_v} [e^{Q^{(n_v)} \int_0^{\tau_v} \alpha_v(t)dt}]_{kj} \ell^{\mathbf{v},\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}} \tag{3.3}$$

where $Q^{(n)} \in \mathbb{R}^{(n+1)\times(n+1)}$ is the transition matrix of the Moran model with n lineages and constant $\alpha = 1$; in particular, $Q^{(n)} = (q^{(n)}_{ij})_{0 \le i,j \le n}$ and

$$q^{(n)}_{ij} = \begin{cases} -i(n-i), & \text{if } i = j, \\ \frac{1}{2} i(n-i), & \text{if } |j-i| = 1, \\ 0, & \text{else.} \end{cases}$$

Finally, in the subsequent three lemmas, we present the formulas to compute $\ell^{\mathbf{v},\mathbf{0}}$ from the partial likelihoods $\ell^{\mathbf{w},\mathbf{t}}$ at the child events $\mathbf{w} \in C_{\mathcal{T}}(\mathbf{v})$.

Lemma 6. *(Split event) Let $\mathbf{v} \in V(\mathcal{T})$ be a split event, where a population w splits into populations v_1, v_2 backwards in time, with each lineage of w moving into v_1 with probability q_1, and into v_2 with probability $q_2 = 1 - q_1$.*

Then $\mathbf{v}$ has a single child event $\mathbf{w} = \mathbf{v} \cup \{w\} \setminus \{v_1, v_2\}$. Let $\boldsymbol{\mu}$ be a configuration of alleles on $\mathbf{v}$, and let $\mu_1 \mathbf{e}_{v_1}, \mu_2 \mathbf{e}_{v_2}, \boldsymbol{\mu}_\cap$ be the subconfigurations on $\{v_1\}, \{v_2\}, \mathbf{v} \cap \mathbf{w}$, respectively, so $\boldsymbol{\mu} = \mu_1 \mathbf{e}_{v_1} + \mu_2 \mathbf{e}_{v_2} + \boldsymbol{\mu}_\cap$. Then

$$\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu}} = \sum_{\mu_w=0}^{n_w} \ell^{\mathbf{w},\tau_w \mathbf{e}_w}_{\boldsymbol{\mu}_\cap + \mu_w \mathbf{e}_w} \sum_{\substack{m_1,m_2: \\ m_1+m_2=n_w}} \binom{n_w}{m_1} q_1^{m_1} q_2^{m_2} \sum_{\substack{j_1,j_2: \\ j_1+j_2=\mu_w}} \frac{\binom{\mu_1}{j_1}\binom{n_{v_1}-\mu_1}{m_1-j_1}}{\binom{n_{v_1}}{m_1}} \frac{\binom{\mu_2}{j_2}\binom{n_{v_2}-\mu_2}{m_2-j_2}}{\binom{n_{v_2}}{m_2}}. \tag{3.4}$$

Lemma 7. *(Join event 1) Let $\mathbf{v} \in V(\mathcal{T})$ be a join event with exactly 1 child event $\mathbf{w}$. In particular, $\mathbf{v}$ is formed when two populations $w_1, w_2 \in \mathbf{w}$ join into an ancestral population v, so $\mathbf{v} = \{v\} \cup \mathbf{w} \setminus \{w_1, w_2\}$.*

Let $\boldsymbol{\mu}_\cap$ be a fixed configuration of alleles on $\mathbf{v} \cap \mathbf{w}$. Define $y^{\boldsymbol{\mu}_\cap} \in \mathbb{R}^{n_{w_1}+n_{w_2}+1}$ and $B \in \mathbb{R}^{(n_{w_1}+n_{w_2}+1)\times(n_v+1)}$ to be the 0-indexed arrays with entries

$$y^{\boldsymbol{\mu}_\cap}_i = \sum_{\substack{j,k: \\ j+k=i}} \frac{\binom{n_{w_1}}{j}\binom{n_{w_2}}{k}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell^{\mathbf{w},\tau_{w_1}\mathbf{e}_{w_1}+\tau_{w_2}\mathbf{e}_{w_2}}_{j\mathbf{e}_{w_1}+k\mathbf{e}_{w_2}+\boldsymbol{\mu}_\cap},$$

$$B_{i,j} = \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}}.$$

Then for $\boldsymbol{\mu} = k\mathbf{e}_v + \boldsymbol{\mu}_\cap$, the partial likelihood is

$$\ell_{\boldsymbol{\mu}}^{\mathbf{v},\mathbf{0}} = [B^+ y^{\boldsymbol{\mu}_\cap}]_k \tag{3.5}$$

with B^+ denoting the Moore-Penrose pseudoinverse of B.

Lemma 8. *(Join event 2) Let $\mathbf{v} \in V(\mathcal{T})$ be a join event with 2 distinct child events $\mathbf{w}_1, \mathbf{w}_2$. In particular, $\mathbf{v}$ is formed when $w_1 \in \mathbf{w}_1$ and $w_2 \in \mathbf{w}_2$ join into an ancestral population v, so $\mathbf{v} = \{v\} \cup \mathbf{w}_1 \cup \mathbf{w}_2 \setminus \{w_1, w_2\}$.*

Let $\boldsymbol{\mu}$ be a configuration of alleles on $\mathbf{v}$, and let $\mu_v \mathbf{e}_v, \boldsymbol{\mu}_{-1}, \boldsymbol{\mu}_{-2}$ be the subconfigurations on $\{v\}, \mathbf{w}_1 \setminus \{w_1\}, \mathbf{w}_2 \setminus \{w_2\}$, respectively, so $\boldsymbol{\mu} = \mu_v \mathbf{e}_v + \boldsymbol{\mu}_{-1} + \boldsymbol{\mu}_{-2}$. Then

$$\ell_{\boldsymbol{\mu}}^{\mathbf{v},\mathbf{0}} = \sum_{\substack{\mu_1,\mu_2: \\ \mu_1+\mu_2=\mu_v}} \frac{\binom{n_{w_1}}{\mu_1}\binom{n_{w_2}}{\mu_2}}{\binom{n_v}{\mu_v}} \ell_{\mu_1\mathbf{e}_{w_1}+\boldsymbol{\mu}_{-1}}^{\mathbf{w}_1,\tau_{w_1}\mathbf{e}_{w_1}} \ell_{\mu_2\mathbf{e}_{w_2}+\boldsymbol{\mu}_{-2}}^{\mathbf{w}_2,\tau_{w_2}\mathbf{e}_{w_2}}. \tag{3.6}$$

3.2.2 Normalizing constant and linear statistics

To compute the probability $\frac{\xi_{\mathbf{x}}}{\|\xi\|_1}$ of a mutation having configuration $\mathbf{x}$, we need not just $\xi_{\mathbf{x}}$, but also the normalizing constant $\|\xi\|_1 = \sum_{\mathbf{x}} \xi_{\mathbf{x}}$ the expected total branch length.

Computing $\|\xi\|_1$ directly is inefficient because of the $O(\prod_{d=1}^{\mathcal{D}} n_d)$ possible entries $\mathbf{x}$. Instead, we can use Algorithm 1 to compute $\|\xi\|_1$, and many more statistics of the SFS, in the same time as $O(1)$ entries $\mathbf{x}$:

Corollary 1. *For $\pi^d \in \mathbb{R}^{n_d+1}, d \in \{1, \ldots, \mathcal{D}\}$, the tensor dot product of the SFS ξ against $\pi^1 \otimes \cdots \otimes \pi^{\mathcal{D}} = [\pi^1_{x_1} \cdots \pi^{\mathcal{D}}_{x_{\mathcal{D}}}]_{x_1,\ldots,x_{\mathcal{D}}}$ is*

$$\begin{aligned}
\xi \odot (\pi^1 \otimes \cdots \otimes \pi^{\mathcal{D}}) &= \sum_{x_1,\ldots,x_{\mathcal{D}}} \xi_{x_1,\ldots,x_{\mathcal{D}}} \pi^1_{x_1} \cdots \pi^{\mathcal{D}}_{x_{\mathcal{D}}} \\
&= \mathrm{DP}(\pi^1, \ldots, \pi^{\mathcal{D}}) \\
&\quad - \left(\prod_{d=1}^{\mathcal{D}} \pi^d_0\right) \mathrm{DP}(\mathbf{e}_0, \ldots, \mathbf{e}_0) - \left(\prod_{d=1}^{\mathcal{D}} \pi^d_{n_d}\right) \mathrm{DP}(\mathbf{e}_{n_1}, \ldots, \mathbf{e}_{n_{\mathcal{D}}}).
\end{aligned}$$

In particular, the total branch length $\|\xi\|_1$ is given by

$$\begin{aligned}
\|\xi\|_1 = \sum_{\mathbf{x}} \xi_{\mathbf{x}} &= \xi \odot (\mathbf{1} \otimes \cdots \otimes \mathbf{1}) \\
&= \mathrm{DP}(\mathbf{1}, \ldots, \mathbf{1}) - \mathrm{DP}(\mathbf{e}_0, \ldots, \mathbf{e}_0) - \mathrm{DP}(\mathbf{e}_{n_1}, \ldots, \mathbf{e}_{n_{\mathcal{D}}})
\end{aligned}$$

with $\mathbf{1}$ the vector with 1 at every coordinate.

Another interesting quantity that is efficiently computed by Corollary 1 is $\mathbb{E}[T^{\mathbf{m}}_{\mathrm{MRCA}}]$, the time of most recent common ancestor for a subsample of size $\mathbf{m} = (m_1, \ldots, m_{\mathcal{D}})$. In

particular, let d' be any leaf population with $m_{d'} > 0$, and for simplicity assume d' is sampled at the present (i.e. d' is not archaic). Then letting $(a)_{b\downarrow} = \prod_{i=0}^{b-1}(a-i)$,

$$\begin{aligned}\mathbb{E}[T^{\mathbf{m}}_{\text{MRCA}}] &= \sum_{x_1,\ldots,x_{\mathcal{D}}} \xi_{x_1,\ldots,x_{\mathcal{D}}} \left(\frac{x_{d'}}{n_{d'}} - \prod_{d=1}^{\mathcal{D}} \frac{(x_d)_{m_d\downarrow}}{(n_d)_{m_d\downarrow}} \right) \\ &= \xi \odot \left(\mathbf{1} \otimes \cdots \otimes \left(\frac{x_{d'}}{n_{d'}} \right)_{x_{d'} \in \{0,\ldots,n_{d'}\}} \otimes \cdots \otimes \mathbf{1} \right) \\ &\quad - \xi \odot \left(\left(\frac{(x_1)_{m_1\downarrow}}{(n_1)_{m_1\downarrow}} \right)_{x_1 \in \{0,\ldots,n_1\}} \otimes \cdots \otimes \left(\frac{(x_{\mathcal{D}})_{m_{\mathcal{D}}\downarrow}}{(n_{\mathcal{D}})_{m_{\mathcal{D}}\downarrow}} \right)_{x_{\mathcal{D}} \in \{0,\ldots,n_{\mathcal{D}}\}} \right).\end{aligned}$$

To see this, note that for a specific individual in d', the mutations above the lineage but below $T^{\mathbf{m}}_{\text{MRCA}}$ are exactly the mutations hitting the lineage (the first term) but *not* hitting all of $\mathbf{m}$ (the second term).

In general, for rank-K tensor $A \in \mathbb{R}^{(n_1+1)\times\cdots\times(n_{\mathcal{D}}+1)}$ with $A = \sum_{k=1}^{K} \mathbf{a}_1^{(k)} \otimes \cdots \otimes \mathbf{a}_{\mathcal{D}}^{(k)}$,

$$\xi \odot A = \sum_{\mathbf{x}} \xi_{\mathbf{x}} A_{\mathbf{x}} = \sum_{k=1}^{K} \xi \odot (\mathbf{a}_1^{(k)} \otimes \cdots \otimes \mathbf{a}_{\mathcal{D}}^{(k)})$$

can be computed in $O(K)$ calls of $\text{DP}(\pi^1, \ldots, \pi^{\mathcal{D}})$, by Corollary 1. We further note that many statistics from population genetics contain terms like $\xi \odot A$, including F_{ST} (Holsinger and Weir, 2009), the "*abba-baba*" D-statistic (Patterson et al., 2012), and many others (e.g. Fay and Wu, 2000; Tajima, 1989), and there has been a longstanding interest in using linear summary statistics of ξ to perform inference (Sainudiin et al., 2011; Durrett, 2008, Ch. 2).

3.3 Computational complexity

Computing $\xi_{\mathbf{x}}$ via Algorithm 1 involves keeping track of the partial likelihoods $\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu},\mathbf{x}}$ at each event $\mathbf{v}$. For a dataset with s unique values of $\mathbf{x}$, the array of partial likelihoods for every $\mathbf{x}, \boldsymbol{\mu}$ at $\mathbf{v}$ is $\ell^{\mathbf{v},\mathbf{0}} = (\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu},\mathbf{x}})_{\boldsymbol{\mu},\mathbf{x}}$ a tensor with $s\prod_{v\in\mathbf{v}}(n_v+1) = O(s\prod_{v\in\mathbf{v}} n_v) = O(sn^{|\mathbf{v}|})$ total entries (where $n = \sum_{d=1}^{\mathcal{D}} n_d$). By contrast, the coalescent approach (Chen, 2012) requires $O(s\prod_{v\in\mathbf{v}} n_v^2) = O(sn^{2|\mathbf{v}|})$ likelihood entries per $\mathbf{v}$ since there are $O(n_v^2)$ states at each v (with $O(n_v)$ states for the number of ancestors, and $O(n_v)$ states for the number of derived alleles). Another alternative, the diffusion approach as implemented in $\partial a \partial i$ (Gutenkunst et al., 2009), numerically integrates the continuous population frequencies forward in time. In particular, if there are D_t populations at time t and the population frequency is discretized into N pieces, then $\partial a \partial i$ has $O(N^{D_t})$ likelihood entries at t, with $N \gg n$ typically. Furthermore, note that $D_t \geq \mathbf{v}$ for each $\mathbf{v}$ at t, since $\mathbf{v}$ is a subset of the populations at t.

We now consider the time cost of each operation in Algorithm 1. We start by considering the "universal constants" that depend only on n_v, and not on the parameters of the

demography. In particular, we compute the matrix exponential of (3.3) as $e^{Q^{(n_v)}\int_0^{\tau_v}\alpha_v(t)dt} = Ue^{\Sigma\int_0^{\tau_v}\alpha_v(t)dt}U^{-1}$ for eigenvalue decomposition $Q^{(n_v)} = U\Sigma U^{-1}$, and the pseudoinverse of (3.5) as $B^+ = V\Sigma^{-1}U^T$ for SVD $B = U\Sigma V^T$. The eigenvalue decomposition and SVD each cost $O(n_v^3)$ but are universal for all demographic parameters. Likewise, the innermost sum of (3.4) is universal for all demographic parameters, but costs $O(n_w^5)$ to compute for all possible values of μ_1, μ_2, μ_w, m_1.

The middle sum of (3.4) is not a universal constant, since it depends on q_1, q_2, but for a fixed demography can be precomputed for all values of μ_1, μ_2, μ_w, and then reused for each SNP $\mathbf{x}$; doing this costs $O(n_w^4)$. Similarly, the truncated SFS $f^v_{n_v}(k)$ in (3.1) can be precomputed for each k, at a total cost of $O(n_v^2)$ (Chapter 2).

Ignoring these precomputation costs, the remaining operations cost $O(s\sum_{\mathbf{v}\in V(\mathcal{T})} n^{|\mathbf{v}|+1})$, since (3.3), (3.4), (3.5), (3.6) can be written to express $\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu},\mathbf{x}}$ as a sum of $O(n)$ terms.

3.4 Application

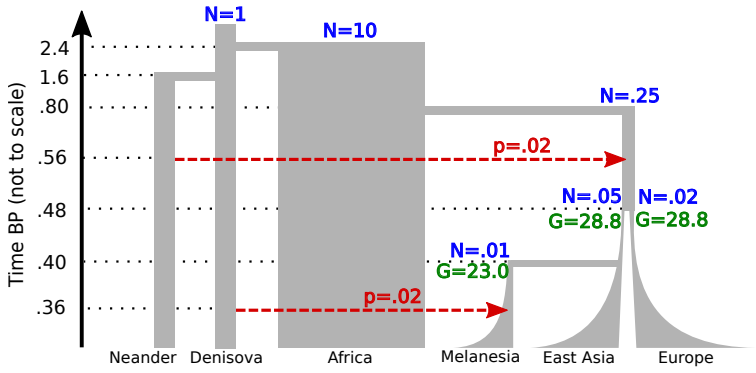


Figure 3.3: A demography loosely based on human history, with 6 leaf populations and 18 demographic parameters. Datasets were simulated with `ms` (Hudson, 2002), and the parameters shown are in `ms`-scaled units. In particular, we simulated 10 datasets, with $n_d = 2$ for Neanderthal and Denisova, and $n_d = 10$ for the remaining populations. Each dataset consisted of 1000 independent loci with $\theta = 10$, with on average 186505.9 SNPs, of which 1516.3 were unique.

Let $\hat{\xi}_{\mathbf{x}}$ denote the observed SFS counts and $\frac{\theta}{2}$ the mutation rate, so that $\mathbb{E}[\hat{\xi}_{\mathbf{x}}] = \frac{\theta}{2}\xi_{\mathbf{x}}$. A typical approach to demographic inference is to maximize the composite likelihood from the Poisson random field approximation,

$$\exp\left(-\frac{\theta}{2}\|\xi\|_1\right)\frac{\left(\frac{\theta}{2}\|\xi\|_1\right)^{\|\hat{\xi}\|_1}}{\|\hat{\xi}\|_1!}\prod_{\mathbf{x}}\left(\frac{\xi_{\mathbf{x}}}{\|\xi\|_1}\right)^{\hat{\xi}_{\mathbf{x}}}. \tag{3.7}$$

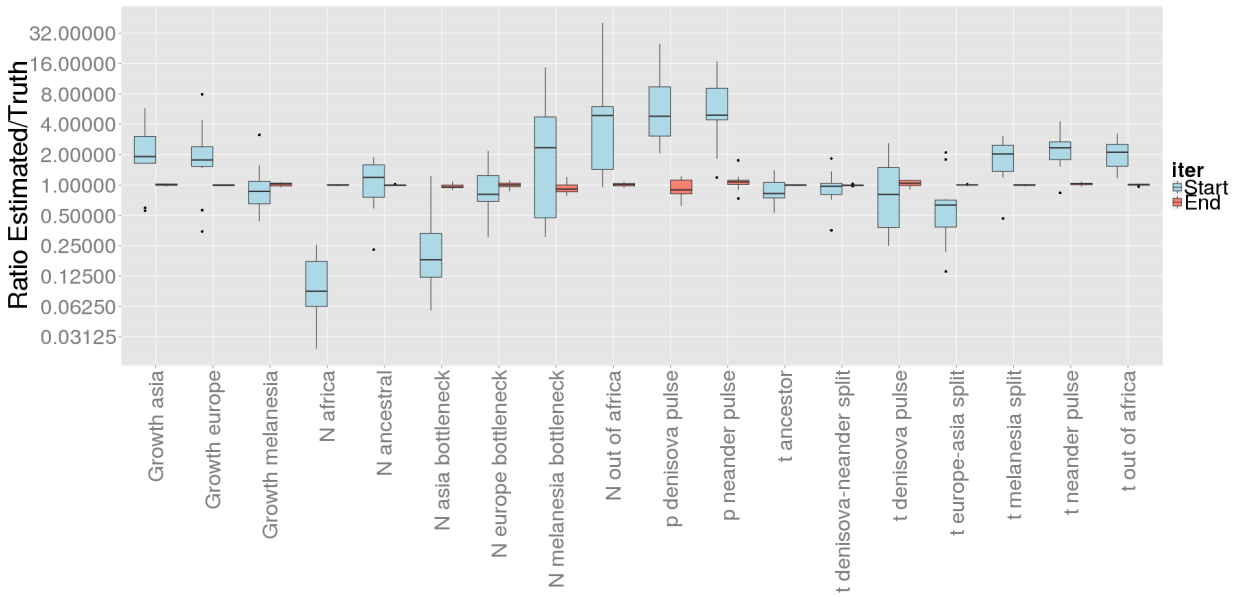


Figure 3.4: For each of the 10 random datasets simulated from Figure 3.3, we chose random initial parameters and searched for a local maximum of the composite likelihood (3.7) with a single run of a conjugate gradient method. The inferred parameters (red boxplots) are very accurate. The average running time of the parameter search (start to finish) was 13.3 hours per dataset.

As an example, we consider a toy demograpy shown in Figure 3.3, with 6 sampled populations and 18 parameters, including admixture events and exponential growth. We simulated 10 random datasets, each with about 186000 SNPs, and used automatic differentiation (Griewank and Corliss, 1991; Bhaskar et al., 2015) with a conjugate gradient method to find a local maximum of the composite likelihood (3.7). The results were highly accurate (Figure 3.4), with each dataset taking an average of 13 hours to complete.

3.5 Proofs

The following two lemmas will be useful for several of the proofs below:

Lemma 9. *For $\mathbf{v} \in V(\mathcal{T})$, the alleles of $\mathcal{M}_{\mathbf{v},\mathbf{t}}$ within $v \in \mathbf{v}$ are exchangeable. That is, the distribution of $\mathcal{M}_{\mathbf{v},\mathbf{t}}$ is invariant to finite permutations of the labels within each $v \in \mathbf{v}$. Furthermore, the labels are independent of the alleles.*

Proof. By construction, none of the lineages at $\mathbf{v}, \mathbf{t}$ are ancestral to each other, i.e. the labels of $\mathcal{M}_{\mathbf{v},\mathbf{t}}$ are unique. Thus the sample genealogy of any finite subsample of $\mathcal{M}_{\mathbf{v},\mathbf{t}}$ is the coalescent, because going backwards in time, coalescence (copying) between each pair

of lineages occurs at rate $\alpha_w(s)$ at vertex w time s. The exchangeability of alleles, and independence of alleles and labels, follows from the coalescent. □

Lemma 10. *Fix event $\mathbf{v} \in V(\mathcal{T})$ and corresponding times $\mathbf{t} = \{t_v : v \in \mathbf{v}\}$. For $u \in \mathbf{v}$, let $\mathcal{I} \subset \{n_u + 1, n_u + 2, \ldots\}$ a (possibly random) collection of indices above n_u, and let $\mu_{t_u}^{u,\mathcal{I}}$ the number of derived alleles in $\{\mathcal{M}_{u,t_u,(i)}\}_{i\in\mathcal{I}}$.*

Then there is conditional independence $\mathbf{x_v} \perp \mu_{t_u}^{u,\mathcal{I}} \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}$.

Proof. Integrate over $\mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}} = \{\mathcal{M}_{v,t_v,(i)}\}_{v\in\mathbf{v},1\le i\le n_v}$ the lowest $\{n_v\}_{v\in\mathbf{v}}$ labeled alleles and $\mathcal{M}_{u,t_u,\mathcal{I}} = \{\mathcal{M}_{u,t_u,(i)}\}_{i\in\mathcal{I}}$ the labeled alleles at $\mathcal{I}$:

$$\begin{aligned}
\mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}, \mu_{t_u}^{u,\mathcal{I}}) &= \mathbb{E}[\mathbb{P}(\mathbf{x_v} \mid \mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}, \mathcal{M}_{u,t_u,\mathcal{I}}) \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}, \mu_{t_u}^{u,\mathcal{I}}] \\
&= \mathbb{E}[\mathbb{P}(\mathbf{x_v} \mid \mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}) \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}, \mu_{t_u}^{u,\mathcal{I}}] \\
&= \mathbb{E}[\mathbb{P}(\mathbf{x_v} \mid \mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}) \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}] \\
&= \mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}})
\end{aligned}$$

with the second equality because higher lineages cannot copy to lower lineages (so $\mathbf{x_v} \perp \mathcal{M}_{u,t_u,\mathcal{I}} \mid \mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}$), and the third equality because of the exchangeability and independence from Lemma 9 (so given $\boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}$, the alleles of $\mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}$ are ordered by a uniform permutation independent of $\mu_{t_u}^{u,\mathcal{I}}$, and the labels of $\mathcal{M}_{\mathbf{v},\mathbf{t},\mathbf{n_v}}$ are independent of $\boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{v}}, \mu_{t_u}^{u,\mathcal{I}}$). □

3.5.1 Proof of Theorem 1

First note that for $d \in \{1, \ldots, \mathcal{D}\}$,

$$\ell^{\{d\},\mathbf{0}} = [\mathbb{P}(\mu_0^d = x_d \mid \mu_0^d = i)]_{0\le i\le n_d} = \mathbf{e}_{x_d}.$$

Then by Lemmas 4,5,6,7,8 and the definition of Algorithm 1, $\mathrm{DP}(\mathbf{e}_{x_1}, \ldots, \mathbf{e}_{x_\mathcal{D}}) = \xi^{\boldsymbol{\rho}} = \xi_{\mathbf{x}}$.

3.5.2 Proof of Lemma 4

$\xi^{\mathbf{v}}$ is the expected number of mutations arising below $\boldsymbol{\tau}_{\mathbf{v}}$ with sampled alleles $\mathbf{x_v}$, per unit mutation rate. $\xi^{\mathbf{v}}$ can be decomposed into 2 parts: the expected number $\xi^{\mathbf{w}}$ arising below $\mathbf{w}, \boldsymbol{\tau}_{\mathbf{w}}$, and the expected number arising within $\mathbf{v} \setminus \mathbf{w}$. That is,

$$\xi^{\mathbf{v}} = \xi^{\mathbf{w}} + \sum_{v\in\mathbf{v}\setminus\mathbf{w}} \sum_{k=1}^{n_v} f_{n_v}^v(k) \ell_{k\mathbf{e}_v}^{\mathbf{v},\mathbf{0}}$$

which proves the first part. For the second part, $\xi_{\mathbf{w}}$ is trivial for $|C_\mathcal{T}(\mathbf{v})| < 2$; otherwise, if $C_\mathcal{T}(\mathbf{v}) = \{\mathbf{w}_1, \mathbf{w}_2\}$,

$$\xi^{\mathbf{w}} = \begin{cases} \xi^{\mathbf{w}_1}, & \text{if } \mathbf{x}_{\mathbf{w}_2} = 0, \\ \xi^{\mathbf{w}_2}, & \text{if } \mathbf{x}_{\mathbf{w}_1} = 0, \\ 0, & \text{else,} \end{cases}$$

and so

$$\xi^{\mathbf{w}} = \sum_{i \neq j} \xi^{\mathbf{w}_i} \mathbb{I}_{\mathbf{x}_{\mathbf{w}_j} = \mathbf{0}} = \sum_{i \neq j} \xi^{\mathbf{w}_i} \prod_{d \in \mathfrak{L}(\mathbf{w}_j)} \mathbb{I}_{x_d = 0} = \sum_{i \neq j} \xi^{\mathbf{w}_i} \prod_{d \in \mathfrak{L}(\mathbf{w}_j)} \ell_{\mathbf{0}}^{\{d\},\mathbf{0}}$$

when $|C_{\mathcal{T}}(\mathbf{v})| = 2$.

3.5.3 Proof of Lemma 5

Define a "quasi-lookdown" Moran model $\mathcal{M}^*$, which is identical to $\mathcal{M}$, except within the n_v lowest lineages of v, where we allow copying in both directions at rate $\frac{\alpha_v(t)}{2}$ (as in the non-lookdown Moran model).

Let $\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}} = \{\boldsymbol{\mu}_{\mathbf{w},\mathbf{s}}\}_{(\mathbf{w},\mathbf{s}) \preceq (\mathbf{v},\mathbf{t})}$ the partial sample path of the allele counts below $\mathbf{v}, \mathbf{t}$, where $(\mathbf{w}, \mathbf{s}) \preceq (\mathbf{v}, \mathbf{t})$ if either $\mathbf{v}$ is an ancestor of $\mathbf{w}$, or $\mathbf{v} = \mathbf{w}$ and $\mathbf{s} \leq \mathbf{t}$ component-wise. It will suffice to show $\mathbb{P}_{\mathcal{M}}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}})$, because then for $\mathbf{t} = \tau_v \mathbf{e}_v + \mathbf{t}_{-v}$,

$$\begin{aligned}
\ell^{\mathbf{v},\mathbf{t}}_{k\mathbf{e}_v + \boldsymbol{\mu}_{-v}} &= \sum_{j=0}^{n_v} \mathbb{P}_{\mathcal{M}}(\boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{t}_{-v}} = j\mathbf{e}_v + \boldsymbol{\mu}_{-v} \mid \boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{t}} = k\mathbf{e}_v + \boldsymbol{\mu}_{-v}) \ell^{\mathbf{v},\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}} \\
&= \sum_{j=0}^{n_v} \mathbb{P}_{\mathcal{M}^*}(\boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{t}_{-v}} = j\mathbf{e}_v + \boldsymbol{\mu}_{-v} \mid \boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{t}} = k\mathbf{e}_v + \boldsymbol{\mu}_{-v}) \ell^{\mathbf{v},\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}} \\
&= \sum_{j=0}^{n_v} \left[e^{M^{(n_v)} \int_0^{\tau_v} \alpha_v(t) dt} \right]_{kj} \ell^{\mathbf{v},\mathbf{t}_{-v}}_{j\mathbf{e}_v + \boldsymbol{\mu}_{-v}}
\end{aligned}$$

as desired.

$\mathbb{P}_{\mathcal{M}}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}})$ follows from a coupling argument. Let $\mathcal{M}_{\preceq \mathbf{v},\mathbf{t}} = \{\mathcal{M}_{\mathbf{w},\mathbf{s}}\}_{(\mathbf{w},\mathbf{s}) \preceq (\mathbf{v},\mathbf{t})}$ the partial sample path below $\mathbf{v}, \mathbf{t}$. We can map the partial sample paths of $\mathcal{M}^*_{\preceq \mathbf{v},\mathbf{t}}$ onto those of $\mathcal{M}_{\preceq \mathbf{v},\mathbf{t}}$ as follows: moving from the bottom to the top of v, whenever a lower label is copied over by a higher label, swap the labels of the lineages above the copying. Then the relabeled sample path has the same distribution as the lookdown construction, since the allele with the higher label is always copied over, and the rate of copying between pairs of lineages is $\alpha_v(t)$. Since this relabeling also leaves μ^v_t unchanged, we have $\mathbb{P}_{\mathcal{M}}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}}) = \mathbb{P}_{\mathcal{M}^*}(\boldsymbol{\mu}_{\preceq \mathbf{v},\mathbf{t}})$.

3.5.4 Proof of Lemma 6

First note that $\mathbf{x}_{\mathbf{v}} = \mathbf{x}_{\mathbf{w}}$ and

$$\begin{aligned}
\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu}} = \sum_{\mu_w = 0}^{n_w} &\mathbb{P}(\mathbf{x}_{\mathbf{w}} \mid \boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{0}} = \boldsymbol{\mu}, \boldsymbol{\mu}^{w}_{\tau_w \mathbf{e}_w} = \boldsymbol{\mu}_{\cap} + \mu_w \mathbf{e}_w) \\
&\times \sum_{\substack{m_1, m_2: \\ m_1 + m_2 = n_w}} \binom{n_w}{m_1} q_1^{m_1} q_2^{m_2} \sum_{\substack{j_1, j_2: \\ j_1 + j_2 = \mu_w}} \frac{\binom{\mu_1}{j_1}\binom{n_{v_1} - \mu_1}{m_1 - j_1}}{\binom{n_{v_1}}{m_1}} \frac{\binom{\mu_2}{j_2}\binom{n_{v_2} - \mu_2}{m_2 - j_2}}{\binom{n_{v_2}}{m_2}}
\end{aligned}$$

by sampling n_w alleles in w from n_{v_1}, n_{v_2} alleles in v_1, v_2, which are exchangeable by Lemma 9.

Next, consider the remaining $n_{v_1}+n_{v_2}-n_w$ alleles of v_1, v_2, and let $\mathcal{I} \subset \{n_w+1, n_w+2, \ldots\}$ be their indices in w. Then because $\mu_{\tau_w}^{w,\mathcal{I}} = \mu_0^{v_1} + \mu_0^{v_2} - \mu_{\tau_w}^{w}$ almost surely and Lemma 10,

$$\mathbb{P}(\mathbf{x_w} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}}, \boldsymbol{\mu}_{\tau_w \mathbf{e}_w}^{\mathbf{w}}) = \mathbb{P}(\mathbf{x_w} \mid \mu_{\tau_w}^{w,\mathcal{I}}, \boldsymbol{\mu}_{\tau_w \mathbf{e}_w}^{\mathbf{w}}) = \mathbb{P}(\mathbf{x_w} \mid \boldsymbol{\mu}_{\tau_w \mathbf{e}_w}^{\mathbf{w}})$$

and so $\mathbb{P}(\mathbf{x_w} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}} = \boldsymbol{\mu}, \boldsymbol{\mu}_{\tau_w \mathbf{e}_w}^{w} = \boldsymbol{\mu}_\cap + \mu_w \mathbf{e}_w) = \ell_{\boldsymbol{\mu}_\cap + \mu_w \mathbf{e}_w}^{\mathbf{w}, \tau_w \mathbf{e}_w}$.

3.5.5 Proof of Lemma 7

First note that

$$\begin{aligned}
y_i^{\boldsymbol{\mu}_\cap} &= \sum_{\substack{j,k: \\ j+k=i}} \frac{\binom{n_{w_1}}{j}\binom{n_{w_2}}{k}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell_{j\mathbf{e}_{w_1} + k\mathbf{e}_{w_2} + \boldsymbol{\mu}_\cap}^{\mathbf{w}, \tau_{w_1}\mathbf{e}_{w_1} + \tau_{w_2}\mathbf{e}_{w_2}} \\
&= \sum_{\substack{j,k: \\ j+k=i}} \mathbb{P}(\boldsymbol{\mu}_{\tau_{w_1}\mathbf{e}_{w_1} + \tau_{w_2}\mathbf{e}_{w_2}}^{\mathbf{w}} = j\mathbf{e}_{w_1} + k\mathbf{e}_{w_2} + \boldsymbol{\mu}_\cap \mid \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i, \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}\cap\mathbf{w}} = \boldsymbol{\mu}_\cap) \\
&\qquad \times \ell_{j\mathbf{e}_{w_1} + k\mathbf{e}_{w_2} + \boldsymbol{\mu}_\cap}^{\mathbf{w}, \tau_{w_1}\mathbf{e}_{w_1} + \tau_{w_2}\mathbf{e}_{w_2}} \\
&= \mathbb{P}(\mathbf{x_v} \mid \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i, \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}\cap\mathbf{w}} = \boldsymbol{\mu}_\cap)
\end{aligned}$$

with the second equality following from exchangeability (Lemma 9) and the third equality from $\mathbf{x_v} = \mathbf{x_w}$.

Next note that

$$\begin{aligned}
&\mathbb{P}(\mathbf{x_v} \mid \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i, \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}\cap\mathbf{w}} = \boldsymbol{\mu}_\cap) \\
=&\sum_{j=0}^{n_v} \mathbb{P}(\boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}} = j\mathbf{e}_v + \boldsymbol{\mu}_\cap \mid \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i, \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}\cap\mathbf{w}} = \boldsymbol{\mu}_\cap) \\
&\qquad \times \mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}} = j\mathbf{e}_v + \boldsymbol{\mu}_\cap, \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i) \\
=&\sum_{j=0}^{n_v} \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}} \mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}} = j\mathbf{e}_v + \boldsymbol{\mu}_\cap, \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} = i)
\end{aligned}$$

with the second equality again due to exchangeability (Lemma 9).

Define $\mathcal{I} \subset \{n_v + 1, n_v + 2, \ldots\}$ so that $\{1, \ldots, n_v\} \cup \mathcal{I}$ are the indices in v of the first n_{w_1}, n_{w_2} alleles in w_1, w_2. Then because $\mu_0^{v,\mathcal{I}} = \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2} - \mu_0^v$ almost surely and Lemma 10,

$$\mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}}, \mu_{\tau_{w_1}}^{w_1} + \mu_{\tau_{w_2}}^{w_2}) = \mathbb{P}(\mathbf{x_v}, \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}}, \mu_0^{v,\mathcal{I}}) = \mathbb{P}(\mathbf{x_v} \mid \boldsymbol{\mu}_{\mathbf{0}}^{\mathbf{v}})$$

and thus

$$y_i^{\boldsymbol{\mu}_\cap} = \sum_{j=0}^{n_v} \frac{\binom{n_v}{j}\binom{n_{w_1}+n_{w_2}-n_v}{i-j}}{\binom{n_{w_1}+n_{w_2}}{i}} \ell_{j\mathbf{e}_v + \boldsymbol{\mu}_\cap}^{\mathbf{v}, \mathbf{0}} = \sum_{j=0}^{n_v} B_{ij} \ell_{j\mathbf{e}_v + \boldsymbol{\mu}_\cap}^{\mathbf{v}, \mathbf{0}}$$

so that for $\boldsymbol{\ell}' = [\ell_{j\mathbf{e}_v + \boldsymbol{\mu}_\cap}^{\mathbf{v}, \mathbf{0}}]_{0 \le j \le n_v} \in \mathbb{R}^{n_v+1}$, we have $y^{\boldsymbol{\mu}_\cap} = B\boldsymbol{\ell}'$ and therefore $\boldsymbol{\ell}' = B^+ y^{\boldsymbol{\mu}_\cap}$.

3.5.6 Proof of Lemma 8

Notice that

$$\begin{aligned}
&\ell^{\mathbf{v},\mathbf{0}}_{\boldsymbol{\mu}_{-1}+\boldsymbol{\mu}_{-2}+\mu_v\mathbf{e}_v} \\
&= \sum_{\substack{\mu_1,\mu_2:\\ \mu_1+\mu_2=\mu_v}} \mathbb{P}(\boldsymbol{\mu}^{\mathbf{w}_1}_{\tau_{w_1}\mathbf{e}_{w_1}} = \boldsymbol{\mu}_{-1}+\mu_1\mathbf{e}_{w_1}, \boldsymbol{\mu}^{\mathbf{w}_2}_{\tau_{w_2}\mathbf{e}_{w_2}} = \boldsymbol{\mu}_{-2}+\mu_2\mathbf{e}_{w_1} \mid \boldsymbol{\mu}^{\mathbf{v}}_{\mathbf{0}} = \boldsymbol{\mu}_{-1}+\boldsymbol{\mu}_{-2}+\mu_v\mathbf{e}_v) \\
&\qquad \times \ell^{\mathbf{w}_1,\tau_{w_1}\mathbf{e}_{w_1}}_{\mu_1\mathbf{e}_{w_1}+\boldsymbol{\mu}_{-1}} \ell^{\mathbf{w}_2,\tau_{w_2}\mathbf{e}_{w_2}}_{\mu_2\mathbf{e}_{w_2}+\boldsymbol{\mu}_{-2}} \\
&= \sum_{\substack{\mu_1,\mu_2:\\ \mu_1+\mu_2=\mu_v}} \frac{\binom{n_{w_1}}{\mu_1}\binom{n_{w_2}}{\mu_2}}{\binom{n_v}{\mu_v}} \ell^{\mathbf{w}_1,\tau_{w_1}\mathbf{e}_{w_1}}_{\mu_1\mathbf{e}_{w_1}+\boldsymbol{\mu}_{-1}} \ell^{\mathbf{w}_2,\tau_{w_2}\mathbf{e}_{w_2}}_{\mu_2\mathbf{e}_{w_2}+\boldsymbol{\mu}_{-2}}
\end{aligned}$$

with the first equality from the Markov property of the Moran process, and the second equality following from the exchangeability of the n_v alleles at vertex v (Lemma 9).

3.5.7 Proof of Corollary 1

Below, we will prove $\mathrm{DP}(\ell^1, \ldots, \ell^{\mathcal{D}})$ is a multilinear function of $\ell^1, \ldots, \ell^{\mathcal{D}}$. The result immediately follows from this, because then

$$\begin{aligned}
\xi \odot (\pi^1 \otimes \cdots \otimes \pi^{\mathcal{D}}) &= \sum_{\mathbf{x}\neq\mathbf{0},\mathbf{n}} \xi_{\mathbf{x}} \pi^1_{x_1} \cdots \pi^{\mathcal{D}}_{x_{\mathcal{D}}} \\
&= \sum_{\mathbf{x}\neq\mathbf{0},\mathbf{n}} \mathrm{DP}(\pi^1_{x_1}\mathbf{e}_{x_1}, \ldots, \pi^{\mathcal{D}}_{x_{\mathcal{D}}}\mathbf{e}_{x_{\mathcal{D}}}) \\
&= \mathrm{DP}(\sum_{x_1=0}^{n_1} \pi^1_{x_1}\mathbf{e}_{x_1}, \ldots, \sum_{x_{\mathcal{D}}=0}^{n_{\mathcal{D}}} \pi^{\mathcal{D}}_{x_{\mathcal{D}}}\mathbf{e}_{x_{\mathcal{D}}}) \\
&\qquad - \mathrm{DP}(\pi^1_0\mathbf{e}_0, \ldots, \pi^{\mathcal{D}}_0\mathbf{e}_0) - \mathrm{DP}(\pi^1_{n_1}\mathbf{e}_{n_1}, \ldots, \pi^{\mathcal{D}}_{n_{\mathcal{D}}}\mathbf{e}_{n_{\mathcal{D}}}) \\
&= \mathrm{DP}(\pi^1, \ldots, \pi^{\mathcal{D}}) - \mathrm{DP}(\pi^1_0\mathbf{e}_0, \ldots, \pi^{\mathcal{D}}_0\mathbf{e}_0) - \mathrm{DP}(\pi^1_{n_1}\mathbf{e}_{n_1}, \ldots, \pi^{\mathcal{D}}_{n_{\mathcal{D}}}\mathbf{e}_{n_{\mathcal{D}}}).
\end{aligned}$$

We now show $\mathrm{DP}(\ell^1, \ldots, \ell^{\mathcal{D}})$ is a multilinear function of $\ell^1, \ldots, \ell^{\mathcal{D}}$. In particular, we note that if event $\mathbf{v}$ has leaf populations $\mathfrak{L}(\mathbf{v}) = (d_1, \ldots, d_{|\mathfrak{L}(\mathbf{v})|})$, then $\ell^{\mathbf{v},\mathbf{t}}$ and $\xi^{\mathbf{v}}$ are multilinear functions of $\ell^{d_1}, \ldots, \ell^{d_{|\mathfrak{L}(\mathbf{v})|}}$. This is trivially true for a leaf $\mathbf{v}, \mathbf{t} = \{d\}, \mathbf{0}$, and otherwise is true by induction, upon noting that (3.1), (3.2) (3.3), (3.4), (3.5), (3.6), express $\ell^{\mathbf{v},\mathbf{t}}$ and $\xi^{\mathbf{v}}$ as sums of multilinear functions of $\ell^{d_1}, \ldots, \ell^{d_{|\mathfrak{L}(\mathbf{v})|}}$. Therefore, the DP is multilinear because it returns $\xi^{\boldsymbol{\rho}}$ a multilinear function of $\ell^1, \ldots, \ell^{\mathcal{D}}$.

Chapter 4

Two Linked Loci with Changing Population Size

The coalescent with recombination (Griffiths and Marjoram, 1997) provides a basic population genetic model for recombination. For a very small number of loci and a constant population size, the likelihood (or sampling probability) can be computed via a recursion (Golding, 1984; Ethier and Griffiths, 1990) or importance sampling (Fearnhead and Donnelly, 2001), allowing for maximum-likelihood and Bayesian estimates of recombination rates (Fearnhead and Donnelly, 2001; Hudson, 2001; McVean et al., 2002; Fearnhead et al., 2004; Fearnhead and Smith, 2005; Fearnhead, 2006).

Jenkins and Song (2009, 2010) recently introduced a new framework based on asymptotic series (in inverse powers of the recombination rate ρ) to approximate the two-locus sampling probability under a constant population size, and developed an algorithm for finding the expansion to an arbitrary order (Jenkins and Song, 2012). They also proved that only a finite number of terms in the expansion is needed to obtain the *exact* two-locus sampling probability as an analytic function of ρ. Bhaskar and Song (2012) partially extended this approach to an arbitrary number of loci and found closed-form formulas for the first two terms in an asymptotic expansion of the multi-locus sampling distribution.

When there are more than a handful of loci, computing the exact sampling probability becomes intractable. A popular and tractable alternative has been to construct composite likelihoods by multiplying the two-locus likelihoods for pairs of SNPs; this pairwise composite likelihood has been used to estimate fine-scale recombination rates in humans (The International HapMap Consortium, 2007; 1000 Genomes Project Consortium, 2010), *Drosophila* (Chan et al., 2012), chimpanzees (Auton et al., 2012), microbes (Johnson and Slatkin, 2009), dogs (Auton et al., 2013), and more, and was used in the discovery of a DNA motif associated with recombination hotspots in some organisms, including humans (Myers et al., 2008), subsequently identified as a binding site of the protein PRDM9 (Myers et al., 2010; Baudat et al., 2010; Berg et al., 2010).

The pairwise composite likelihood was first suggested by Hudson (2001) for an infinite-sites model. The software package LDhat (McVean et al., 2004; Auton and McVean, 2007)

implemented the pairwise composite likelihood for a finite-sites model, and embedded it within a Bayesian MCMC algorithm for inference. Chan et al. (2012) modified this algorithm in their program LDhelmet to efficiently utilize aforementioned asymptotic formulas for the sampling probability, among other improvements. The program LDhot (Myers et al., 2005; Auton et al., 2014) uses the composite likelihood as a test statistic to detect recombination hotspots, in conjunction with coalescent simulation to determine appropriate null distributions.

Because of mathematical and computational challenges, LDhat, LDhelmet, and LDhot all assume a constant population size model to compute the two-locus sampling probabilities. This is an unrealistic assumption, and it would be desirable to account for known demographic events, such as bottlenecks or population growth. Furthermore, Johnston and Cutler (2012) observed that a sharp bottleneck, followed by rapid growth, can lead LDhat to infer many spurious recombination hotspots, possibly due to the incorrect assumption of a constant population size history.

In this chapter, we show how to compute the two-locus sampling probability under variable population size histories that are piecewise constant. We develop two distinct methods for this task. Both approaches rely heavily on the Moran model.

The first approach is an exact formula, introduced in Theorem 2, that involves exponentiating sparse m-by-m matrices containing $O(m)$ nonzero entries, where $m = O(n^6)$, with n being the sample size. We derive this formula by constructing a modification of the standard two-locus Moran process (which we call an "augmented Moran model"), in which sample paths can be coupled with the two-locus coalescent, and by applying a reversibility argument.

The second approach is a highly efficient importance sampler, based on an optimal proposal distribution that we characterize in Theorem 3. Here, we directly use the standard two-locus Moran model to approximate certain terms in the optimal proposal distribution. Theorem 3 generalizes previous results for the constant size case, which have been used to construct importance samplers for both the single-population, two-locus case (Fearnhead and Donnelly, 2001) and for other time-homogeneous coalescent scenarios (Stephens and Donnelly, 2000; De Iorio and Griffiths, 2004; Griffiths et al., 2008; Koskela et al., 2015). The key ideas of Theorem 3 should similarly generalize to other contexts of importance sampling a time-inhomogeneous coalescent.

The importance sampler has the disadvantage of being a Monte Carlo method. However, it has advantages in certain contexts. It is easily parallelizable, and techniques such as bridge sampling (Meng and Wong, 1996; Fearnhead and Donnelly, 2001) could yield further computational savings. The importance sampler yields a posterior sample of two-locus ARGs, which may be used to obtain interesting genealogical information. In addition, the importance sampler does not require the user to compute the sampling probability of every possible two-locus dataset, in contrast to Theorem 2. This is useful if the user only needs to consider a subset of potential datasets.

We examine the runtime and accuracy of our two approaches. We empirically show our importance sampler to be extremely efficient, with an average effective sample size of almost

98%. We then apply the exact algorithm to study the effect of a sharp population bottleneck (followed by a rapid population expansion) on linkage disequilibrium, and note that the expected r^2, a statistic commonly used to detect linkage disequilibrium, has less power under this demographic scenario. We also study to what extent using the correct two-locus sampling probabilities in the composite likelihood, as opposed to assuming an unrealistic constant population size history, improves the estimation of fine-scale recombination rates.

4.1 Background

Here we describe our notational convention, and review some key concepts regarding the coalescent with recombination and the two-locus Moran model.

Note that the notation in this chapter differs substantially from the previous two chapters, which dealt purely with the one-locus case, but included multiple populations. By contrast, here we consider the sampling probabilities at two loci, but in just a single population.

4.1.1 Notation

Let $\frac{\theta}{2}$ denote the mutation rate per locus per unit time, $\mathbf{P} = (P_{ij})_{i,j\in\mathcal{A}}$ the transition probabilities between alleles given a mutation, and $\mathcal{A}$ the set of alleles. Let $\frac{\rho}{2}$ denote the recombination rate per unit time. We consider a single panmictic population, with piecewise-constant effective population sizes. In particular, we assume D pieces, with endpoints $-\infty = t_{-D} < t_{-D+1} < \cdots < t_{-1} < t_0 = 0$, where 0 corresponds to the present and $t < 0$ corresponds to a time in the past. The piece $(t_d, t_{d+1}]$ is assumed to have scaled population size $\frac{1}{\alpha_d}$. Going backwards in time, two lineages coalesce (find a common ancestor) at rate α_d within the interval $(t_d, t_{d-1}]$.

We allow the haplotypes to have missing (unobserved) alleles at each locus, and use $*$ to denote such alleles. We denote each haplotype as having type a, b, or c, where a haplotypes are only observed at the left locus, b haplotypes are only observed at the right locus, and c haplotypes are observed at both loci. We use $\mathbf{n} = \{n_{ij}, n_{i*}, n_{*j}\}_{i,j\in\mathcal{A}}$ to denote the configuration of an unordered collection of two-locus haplotypes, with n_{ij} corresponding to the number of haplotypes with allele i at the first locus and allele j at the second locus, and so on.

Suppose $\mathbf{n}$ has $n^{(abc)} = (n^{(a)}, n^{(b)}, n^{(c)})$ haplotypes of type a, b, c respectively. We define the *sampling probability* $\mathbb{P}_t(\mathbf{n})$ to be the probability of sampling $\mathbf{n}$ at time t, given that we observed $n^{(a)}, n^{(b)}, n^{(c)}$ haplotypes of type a, b, c, under the coalescent with recombination (next subsection).

4.1.2 The ARG and the coalescent with recombination

The Ancestral Recombination Graph (ARG) is the multi-locus genealogy relating a sample (Figure 4.1). The coalescent with recombination (Griffiths, 1991) gives the limiting distri-

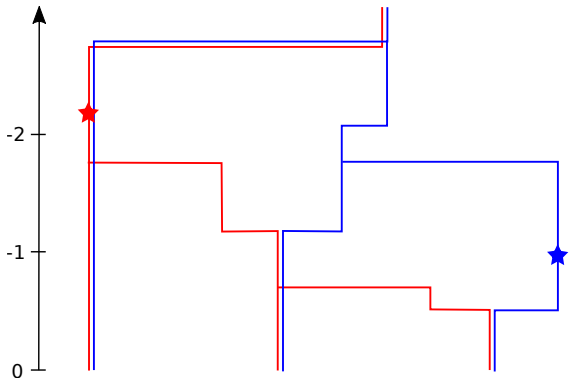


Figure 4.1: An ARG at two loci with $n = 3$. Coalescence is when two lineages find a common parent. Recombination is when a lineage inherits from two parents. Mutations are shown as stars.

End state	Rate
$n_t^{(a)}+1, n_t^{(b)}+1, n_t^{(c)}-1$	$\frac{\rho}{2} n_t^{(c)}$
$n_t^{(a)}-1, n_t^{(b)}, n_t^{(c)}$	$\alpha_d n_t^{(a)} \left(\frac{n_t^{(a)}-1}{2} + n_t^{(c)}\right)$
$n_t^{(a)}, n_t^{(b)}-1, n_t^{(c)}$	$\alpha_d n_t^{(b)} \left(\frac{n_t^{(b)}-1}{2} + n_t^{(c)}\right)$
$n_t^{(a)}, n_t^{(b)}, n_t^{(c)}-1$	$\alpha_d \binom{n_t^{(c)}}{2}$
$n_t^{(a)}-1, n_t^{(b)}-1, n_t^{(c)}+1$	$\alpha_d n_t^{(a)} n_t^{(b)}$

Table 4.1: Backward in time transition rates of $n_t^{(abc)} = (n_t^{(a)}, n_t^{(b)}, n_t^{(c)})$ within time interval $(t_d, t_{d+1}]$ under the coalescent with recombination.

bution of the ARG under a wide class of population models, including the Wright-Fisher model and the Moran model.

Let $n_t^{(c)}$ be the number of lineages at time t that are ancestral to the observed present-day sample at both loci. Similarly, let $n_t^{(a)}$ and $n_t^{(b)}$ be the number of lineages that are ancestral at only the a or b locus, respectively. Under the coalescent with recombination, $n_t^{(abc)} = (n_t^{(a)}, n_t^{(b)}, n_t^{(c)})$ is a backwards in time Markov chain, where each c type lineage splits (recombines) into one a and one b lineage at rate $\frac{\rho}{2}$, and each pair of lineages coalesces at rate α_d within the time interval $(t_d, t_{d+1}]$. Table 4.1 gives the transition rates of $n_t^{(abc)}$.

After sampling the history of coalescence and recombination events $\{n_t^{(abc)}\}_{t \le 0}$, we drop mutations down at rate $\frac{\theta}{2}$ per locus, with alleles mutating according to $\mathbf{P}$, and the alleles of the common ancestor assumed to be at the stationary distribution. This gives us a sample path $\{\mathbf{n}_t\}_{t \le 0}$, where $\mathbf{n}_0$ is the observed sample at the present, and $\mathbf{n}_t$ is the collection of ancestral haplotypes at time t. Under this notation, the sampling probability at time t is defined as

$$\mathbb{P}_t(\mathbf{n}) \quad := \quad \mathbb{P}(\mathbf{n}_t = \mathbf{n} \mid n_t^{(abc)} = n^{(abc)}). \tag{4.1}$$

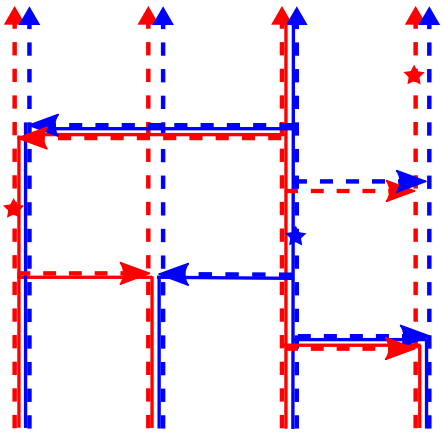


Figure 4.2: Moran model at locus a (red dashes) and locus b (blue dashes), with mutations (stars). The embedded ARG is shown in thin solid lines. Note the ARG has a simultaneous recombination and coalescence, which has probability 0 under the coalescent.

4.1.3 Two-locus Moran model

We review the Moran model with recombination, as described by Ethier and Kurtz (1993); Donnelly and Kurtz (1999). This process is directly used in the proposal distribution of our importance sampler.

The Moran model with N lineages is a finite population model evolving forward in time. In particular, let $\mathbf{M}_t$ denote a collection of N two-locus haplotypes at time t (with no missing alleles). Then $\mathbf{M}_t$ is a Markov chain going forwards in time that changes due to mutation, recombination, and copying events.

Let $\mathbf{\Lambda}^d_{(N)}$ denote the transition matrix of $\mathbf{M}_t$ within $(t_d, t_{d+1}]$. We describe the rates of $\mathbf{\Lambda}^d_{(N)}$. For the mutation events, each allele mutates at rate $\frac{\theta}{2}$ according to transition matrix $\mathbf{P}$. For the copying events, each lineage of $\mathbf{M}_t$ copies its haplotype onto each other lineage at rate $\frac{\alpha_d}{2}$ within the time interval $(t_d, t_{d+1}]$. Finally, for recombination, each lineage of $\mathbf{M}_t$ experiences a recombination event at rate $\frac{\rho}{2}$, at which point it copies the left allele from a uniformly chosen lineage of $\mathbf{M}_t$, and independently copies the right allele from another uniformly chosen lineage of $\mathbf{M}_t$. See Figure 4.2 for illustration.

$\mathbf{M}_{-\infty}$ is sampled by drawing from the stationary distribution $\boldsymbol{\lambda}^{-D}_{(N)}$ of $\mathbf{\Lambda}^{-D}_{(N)}$. The probability of $\mathbf{M}_t$ is then given by

$$[\mathbb{P}_{(N)}(\mathbf{M}_t = \mathbf{M})]_{\mathbf{M}} \quad = \quad \boldsymbol{\lambda}^{-D}_{(N)} \prod_{d=-D+1}^{-1} e^{\mathbf{\Lambda}^d_{(N)}[\min(t,t_{d+1})-\min(t,t_d)]}, \tag{4.2}$$

where $[\mathbb{P}_{(N)}(\mathbf{M}_t = \mathbf{M})]_{\mathbf{M}}$ and $\boldsymbol{\lambda}^{-D}_{(N)}$ are row vectors here. Let $\mathbb{P}(\mathbf{n} \mid \mathbf{M})$ denote the probability of sampling $\mathbf{n}$ by drawing haplotypes without replacement from $\mathbf{M}$. Then the likelihood of

observing $\mathbf{n}$ at time t under the Moran model is

$$\mathbb{P}_t^{(N)}(\mathbf{n}) \;=\; \sum_{\mathbf{M}} \mathbb{P}_{(N)}(\mathbf{M}_t = \mathbf{M})\mathbb{P}(\mathbf{n} \mid \mathbf{M}). \tag{4.3}$$

As $N \to \infty$, $\mathbb{P}_t^{(N)}(\mathbf{n}) \to \mathbb{P}_t(\mathbf{n})$, i.e. the Moran likelihood converges to the likelihood of the coalescent with recombination (Donnelly and Kurtz, 1999). However, for $N < \infty$, we generally have $\mathbb{P}_t^{(N)}(\mathbf{n}) \neq \mathbb{P}_t(\mathbf{n})$. This is because the embedded ARG within the two-locus Moran model can have simultaneous recombination and coalescent events (Figure 4.2), which has probability 0 under the coalescent. This contrasts with the 1-locus case: in the 1-locus Moran model, the embedded genealogy has the exact same distribution as the coalescent, for all N.

4.2 Theoretical Results

In this section, we describe our main theoretical results. All proofs are deferred to Section 4.5.

4.2.1 Augmented two-locus Moran model

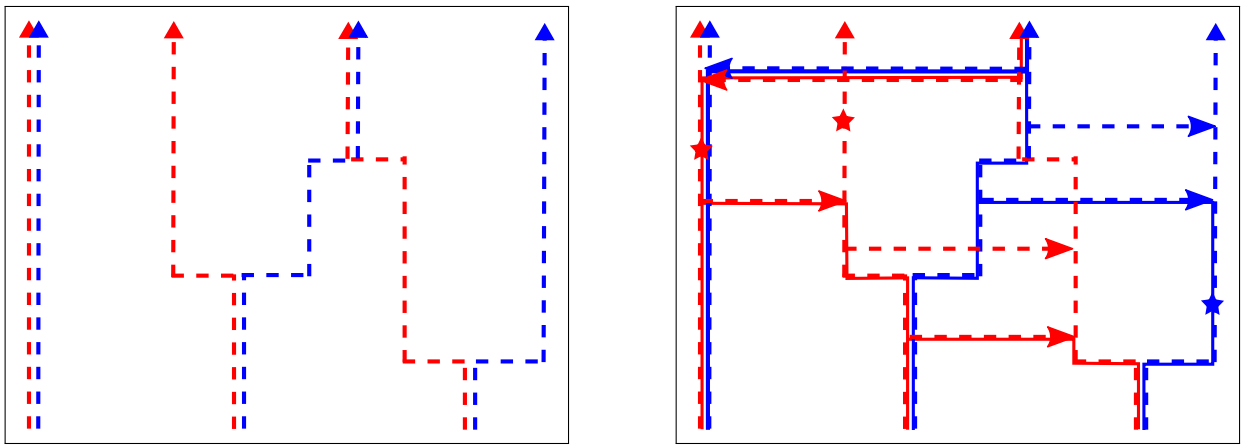
To obtain our main theoretical result on sampling probability (see Theorem 2), we introduce an "augmented two-locus Moran model", denoted $\tilde{\mathbf{M}}_t^*$. At a single locus, this process is exactly the same as the usual Moran model $\mathbf{M}_t$. However, at 2 loci, $\tilde{\mathbf{M}}_t^*$ has an embedded ARG whose distribution agrees with the two-locus coalescent.

We define $\tilde{\mathbf{M}}_t^*$ in detail in Section 4.5.1. Briefly, the biggest difference between the "original" and "augmented" Moran models is that all lineages in $\mathbf{M}_t$ are fully-specified c-types, while $\tilde{\mathbf{M}}_t^*$ allows partially specified a- and b-types. To generate $\tilde{\mathbf{M}}_t^*$, we start by going *backwards* in time, dropping recombination events (c-types splitting into a- and b-types) and "cross-coalescence" events (coalescence of (a, b)-pairs into c-types). After generating these recombinations and cross-coalescences, we go *forwards* in time, dropping mutation and copying events to obtain the sample $\mathbf{n}$ at the present. We illustrate this process in Figure 4.3.

Working with $\tilde{\mathbf{M}}_t^*$ directly is inconvenient, due to some events happening forwards-in-time, and others happening backwards-in-time. We thus use a different Moran-like process $\tilde{\mathbf{M}}_t$ that is entirely forwards-in-time, with rates $\tilde{\mathbf{\Lambda}}^d$ given in Section 4.2.2. While $\tilde{\mathbf{M}}_t$ does not have the same sampling probabilities as the coalescent, we use a reversibility argument to relate its distribution to $\tilde{\mathbf{M}}_t^*$, to derive Theorem 2.

4.2.2 A formula for the sampling probability

Let $\mathcal{N} = \{\mathbf{n} : n^{(abc)} = (k, k, n-k), 0 \leq k \leq n\}$ denote the collection of sample configurations with n specified alleles at each locus. We refer to n as the size of $\mathbf{n}$. For each interval $(t_d, t_{d+1}]$, let $\tilde{\mathbf{\Lambda}}^d$ be a square matrix indexed by $\mathcal{N}$, with entries given in Table 4.2. Let $\mathbf{\Gamma}^d$ be a tridiagonal square matrix indexed by $\{0, 1, \ldots, n\}$, with $\Gamma^d_{m,m-1} = \frac{\rho}{2}m$, $\Gamma^d_{m,m+1} = (n-m)^2\alpha_d$,



(a) (b)

Figure 4.3: Augmented two-locus Moran model. (a) First, backwards in time, add recombinations (c-type $\to$ a-type and b-type) and "cross-coalescences" (a-type and b-type $\to$ c-type). (b) Next, forward in time, add copying and mutation events. The embedded ARG is shown in thin solid lines.

$\mathbf{m}$	$\tilde{\mathbf{\Lambda}}^d_{\mathbf{n},\mathbf{m}}$
$\mathbf{n}-\mathbf{e}_{i*}+\mathbf{e}_{j*}$	$\alpha_d n_{i*}(\frac{1}{2}n_{j*}+\sum_{k\in\mathcal{A}} n_{jk})+\frac{\theta}{2}P_{ij}n_{i*}$
$\mathbf{n}-\mathbf{e}_{*i}+\mathbf{e}_{*j}$	$\alpha_d n_{*i}(\frac{1}{2}n_{*j}+\sum_{k\in\mathcal{A}} n_{kj})+\frac{\theta}{2}P_{ij}n_{*i}$
$\mathbf{n}-\mathbf{e}_{ij}+\mathbf{e}_{kl}$	$\frac{\alpha_d}{2}n_{ij}n_{kl}+\frac{\theta}{2}(\delta_{ik}P_{jl}+\delta_{jl}P_{ik})n_{ij}$
$\mathbf{n}-\mathbf{e}_{ij}+\mathbf{e}_{i*}+\mathbf{e}_{*j}$	$\frac{\rho}{2}n_{ij}$
$\mathbf{n}-\mathbf{e}_{i*}-\mathbf{e}_{*j}+\mathbf{e}_{ij}$	$n_{i*}n_{*j}\alpha_d$
$\mathbf{n}$	$-\alpha_d\binom{n^{(a)}+n^{(b)}+n^{(c)}}{2}-\frac{\rho}{2}n^{(c)}-\frac{\theta}{2}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{A}\cup\{*\}}(n_{ij}+n_{ji})$

Table 4.2: Nonzero entries of the rate matrix $\tilde{\mathbf{\Lambda}}^d$ for the interval $(t_d, t_{d+1}]$.

and $\Gamma^d_{m,m}=-\Gamma^d_{m,m-1}-\Gamma^d_{m,m+1}$. Then we have the following result, a proof of which is provided in Section 4.5.1:

Theorem 2. *Let $(\gamma^d_0,\ldots,\gamma^d_n)$ be the stationary distribution of $\mathbf{\Gamma}^d$, and let the row vector $\tilde{\boldsymbol{\gamma}}^d$ be indexed by $\mathcal{N}$, with $\tilde{\gamma}^d_{\mathbf{n}}=\gamma^d_m$ if $\mathbf{n}$ has m lineages of type c. Denote the stationary distribution of $\tilde{\mathbf{\Lambda}}^d$ by the row vector $\tilde{\boldsymbol{\lambda}}^d=(\lambda^d_{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$. Let $\odot$ and $\div$ denote component-wise multiplication and division, and recursively define the row vector $\mathbf{p}^d=(p^d_{\mathbf{n}})_{\mathbf{n}\in\mathcal{N}}$ by*

$$\begin{aligned}\mathbf{p}^{-D+1}&=\tilde{\boldsymbol{\lambda}}^{-D}\div\tilde{\boldsymbol{\gamma}}^{-D}\\ \mathbf{p}^{d+1}&=\left[(\mathbf{p}^d\odot\tilde{\boldsymbol{\gamma}}^d)e^{\tilde{\mathbf{\Lambda}}^d(t_{d+1}-t_d)}\right]\div\tilde{\boldsymbol{\gamma}}^d.\end{aligned}\tag{4.4}$$

Then, for $\mathbf{n}\in\mathcal{N}$, we have $\mathbb{P}_0(\mathbf{n})=p^0_{\mathbf{n}}$.

Note that Theorem 2 gives $\mathbb{P}_0(\mathbf{n})$ for $\mathbf{n}\in\mathcal{N}$. This includes all fully specified $\mathbf{n}$, i.e. with $n^{(abc)}=(0,0,n)$, and suffices for the application considered in Section 4.4.2. If neces-

sary, $\mathbb{P}_0(\mathbf{n})$ for partially specified $\mathbf{n}$ can be computed by summing over the fully specified configurations and sampling without replacement.

For $|\mathcal{A}| = 2$, $\tilde{\mathbf{\Lambda}}^d$ is an $O(n^6) \times O(n^6)$ matrix, so naively computing the matrix multiplication in Theorem 2 would cost $O(n^{12})$ time. However, $\tilde{\mathbf{\Lambda}}^d$ is sparse, with $O(n^6)$ nonzero entries, allowing more efficient algorithms to compute Theorem 2 to numerical precision in $O(n^6\mathcal{T})$, where $\mathcal{T}$ is some finite number of matrix-vector multiplications. In Section 4.3.2, we discuss the computational complexity in more detail, and also discuss how to numerically compute the matrix exponential.

4.2.3 Importance sampling

We now describe an alternative method for computing $\mathbb{P}_0(\mathbf{n})$ via importance sampling on the sample paths $\mathbf{n}_{\le 0} = \{\mathbf{n}_t\}_{t\le 0}$.

Let the proposal distribution $Q(\mathbf{n}_{\le 0})$ be a probability distribution on $\{\mathbf{n}_{\le 0} : \mathbf{n}_0 = \mathbf{n}\}$ whose support contains that of $\mathbb{P}(\mathbf{n}_{\le 0} \mid \mathbf{n}_0 = \mathbf{n})$. Then we have

$$\mathbb{P}_0(\mathbf{n}) \;=\; \int_{\mathbf{n}_{\le 0}:\mathbf{n}_0=\mathbf{n}} \frac{d\mathbb{P}(\mathbf{n}_{\le 0})}{dQ(\mathbf{n}_{\le 0})} dQ(\mathbf{n}_{\le 0}),$$

and so, if $\mathbf{n}_{\le 0}^{(1)}, \ldots, \mathbf{n}_{\le 0}^{(K)} \sim Q$ i.i.d., the sum

$$\frac{1}{K}\sum_{k=1}^{K} \frac{d\mathbb{P}(\mathbf{n}_{\le 0}^{(k)})}{dQ(\mathbf{n}_{\le 0}^{(k)})} \tag{4.5}$$

converges almost surely to $\mathbb{P}_0(\mathbf{n})$ as $K \to \infty$ by the Law of Large Numbers. Hence, (4.5) provides a Monte Carlo approximation to $\mathbb{P}_0(\mathbf{n})$. The optimal proposal is the posterior distribution $Q_{\mathrm{opt}}(\mathbf{n}_{\le 0}) = \mathbb{P}(\mathbf{n}_{\le 0} \mid \mathbf{n}_0)$, for then (4.5) is exactly

$$\frac{1}{K}\sum_{k=1}^{K} \frac{d\mathbb{P}(\mathbf{n}_{\le 0}^{(k)})}{d\mathbb{P}(\mathbf{n}_{\le 0}^{(k)} \mid \mathbf{n}_0)} = \frac{1}{K}\sum_{k=1}^{K} \frac{d\mathbb{P}(\mathbf{n}_{\le 0}^{(k)})}{d\mathbb{P}(\mathbf{n}_{\le 0}^{(k)})/\mathbb{P}(\mathbf{n}_0)} = \mathbb{P}(\mathbf{n}_0),$$

even for $K = 1$.

The following theorem, which we prove in Section 4.5.2, characterizes the optimal posterior distribution $Q_{\mathrm{opt}}(\mathbf{n}_{\le 0}) = \mathbb{P}(\mathbf{n}_{\le 0} \mid \mathbf{n}_0)$ for variable population size:

Theorem 3. *The process $\{\mathbf{n}_t\}_{t\le 0}$ is a backward-in-time Markov chain with inhomogeneous rates, whose rate matrix at time t is given by*

$$q_{\mathbf{n},\mathbf{m}}^{(t)} \;=\; \begin{cases} \phi_{\mathbf{n},\mathbf{m}}^{(t)} \frac{\mathbb{P}_t(\mathbf{m})}{\mathbb{P}_t(\mathbf{n})}, & \text{if } \mathbf{m} \neq \mathbf{n}, \\ \phi_{\mathbf{n},\mathbf{n}}^{(t)} - \frac{d}{dt}\log \mathbb{P}_t(\mathbf{n}), & \text{if } \mathbf{m} = \mathbf{n}, \end{cases}$$

where $\boldsymbol{\phi}^{(t)} = (\phi_{\mathbf{n},\mathbf{m}}^{(t)})$ is a square matrix, indexed by configurations $\mathbf{n}$, with entries given by Table 4.3 and equal to

$$\phi_{\mathbf{n},\mathbf{m}}^{(t)} \;=\; \frac{d}{ds}\Big[\mathbb{P}(n_{t-s}^{(abc)} = m^{(abc)} \mid n_t^{(abc)} = n^{(abc)})\,\mathbb{P}(\mathbf{n}_t = \mathbf{n} \mid \mathbf{n}_{t-s} = \mathbf{m}, n_t^{(abc)} = n^{(abc)})\Big]\Big|_{s=0} \tag{4.6}$$

$\mathbf{m}$	$\phi^{(t)}_{\mathbf{n},\mathbf{m}}$
$\mathbf{n}-\mathbf{e}_{i*}+\mathbf{e}_{j*}$	$\frac{\theta}{2}P_{ji}(n_{j*}+1)$
$\mathbf{n}-\mathbf{e}_{*i}+\mathbf{e}_{*j}$	$\frac{\theta}{2}P_{ji}(n_{*j}+1)$
$\mathbf{n}-\mathbf{e}_{ij}+\mathbf{e}_{kl}$	$\frac{\theta}{2}(\delta_{ik}P_{lj}+\delta_{jl}P_{ki})(n_{kl}+1)$
$\mathbf{n}-\mathbf{e}_{ij}+\mathbf{e}_{i*}+\mathbf{e}_{*j}$	$\frac{\rho}{2}n^{(c)}(n_{i*}+1)(n_{*j}+1)$
$\mathbf{n}-\mathbf{e}_{ij}$	$\alpha_d\binom{n^{(c)}}{2}(n_{ij}-1)$
$\mathbf{n}-\mathbf{e}_{i*}$	$\alpha_d\big[\binom{n^{(a)}}{2}(n_{i*}-1)+n^{(a)}n^{(c)}\sum_j n_{ij}\big]$
$\mathbf{n}-\mathbf{e}_{*i}$	$\alpha_d\big[\binom{n^{(b)}}{2}(n_{*i}-1)+n^{(b)}n^{(c)}\sum_j n_{ji}\big]$
$\mathbf{n}$	$-\alpha_d\binom{n}{2}-\frac{\rho}{2}n^{(c)}-\frac{\theta}{2}\sum_i(1-P_{ii})[n_{i*}+n_{*i}+\sum_j(n_{ij}+n_{ji})]$

Table 4.3: Nonzero entries of the $\boldsymbol{\phi}^{(t)}$ matrix of Theorem 3, for $t \in (t_d, t_{d+1}]$.

Theorem 3 generalizes previous results for the optimal proposal distribution in the constant size case (Stephens and Donnelly, 2000; Fearnhead and Donnelly, 2001). In that case, the conditional probability of the parent $\mathbf{m}$ of $\mathbf{n}$ is $\phi_{\mathbf{n},\mathbf{m}}\frac{\mathbb{P}(\mathbf{m})}{\mathbb{P}(\mathbf{n})}$. Note the constant size case is time-homogeneous, so the dependence on t is dropped, and the waiting times between events in the ARG are not sampled (i.e., only the embedded jump chain of $\mathbf{n}_{\leq 0}$ is sampled).

We construct our proposal distribution $\hat{Q}(\mathbf{n}_{\leq 0})$ by approximating the optimal proposal distribution $Q_{\text{opt}}(\mathbf{n}_{\leq 0}) = \mathbb{P}(\mathbf{n}_{\leq 0} \mid \mathbf{n}_0)$. We start by choosing a grid of points $-\infty < \tau_1 < \tau_2 < \cdots < \tau_J = 0$, then set $\hat{Q}$ to be a backwards in time Markov chain, whose rates at $t \in (\tau_j, \tau_{j+1})$ are the linear interpolation

$$\hat{q}^{(t)}_{\mathbf{n},\mathbf{m}} = \frac{\tau_{j+1}-t}{\tau_{j+1}-\tau_j}\hat{q}^{(\tau_j)}_{\mathbf{n},\mathbf{m}} + \frac{t-\tau_j}{\tau_{j+1}-\tau_j}\hat{q}^{(\tau_{j+1})}_{\mathbf{n},\mathbf{m}}, \tag{4.7}$$

with the rates at the grid points given by

$$\hat{q}^{(\tau_j)}_{\mathbf{n},\mathbf{m}} = \begin{cases} \phi^{(\tau_j)}_{\mathbf{n},\mathbf{m}}\frac{\hat{\mathbb{P}}_{\tau_j}(\mathbf{m})}{\hat{\mathbb{P}}_{\tau_j}(\mathbf{n})}, & \text{if } \mathbf{m} \neq \mathbf{n}, \\ -\sum_{\boldsymbol{\nu}\neq\mathbf{n}}\hat{q}^{(\tau_j)}_{\mathbf{n},\boldsymbol{\nu}}, & \text{if } \mathbf{m} = \mathbf{n}, \end{cases}$$

with $\hat{\mathbb{P}}_{\tau_j}(\mathbf{n})$ an approximation to the likelihood $\mathbb{P}_{\tau_j}(\mathbf{n})$. In particular, we set $\hat{\mathbb{P}}_{\tau_j}(\mathbf{n}) = \mathbb{P}^{(N)}_{\tau_j}(\mathbf{n})$, the likelihood for the standard (not augmented) two-locus Moran process, defined in Section 4.1.3.

To sample from $\hat{Q}$, we note that for configuration $\mathbf{n}$ at time t, the time $S < t$ of the next event has CDF $\mathbb{P}(S < s) = \exp(\int_S^t \hat{q}^{(u)}_{\mathbf{n},\mathbf{n}}du)$ for $s < t$. Thus, S can be sampled by first sampling $X \sim \text{Uniform}(0,1)$, and then solving for $\log(X) = \int_S^t \hat{q}^{(u)}_{\mathbf{n},\mathbf{n}}du$ via the quadratic formula (since $\hat{q}^{(u)}_{\mathbf{n},\mathbf{n}}$ is piecewise linear; see (4.7)). Having sampled S, we can then sample the next configuration $\mathbf{m}$ with probability $-\hat{q}^{(S)}_{\mathbf{n},\mathbf{m}}/\hat{q}^{(S)}_{\mathbf{n},\mathbf{n}}$.

As detailed in Section 4.3.3, $\hat{Q}$ is a highly efficient proposal distribution, yielding an average effective sample size (ESS) of almost 98% per sample for the demography and ρ values we considered.

4.3 Runtime, Accuracy, and Efficiency Results

We now discuss the computational complexity and empirical runtime of our methods. In doing so, we also develop some insight into the computational details of our algorithms. We start by discussing how to efficiently multiply a vector against the exponential of a sparse matrix. This technique is needed for both Theorem 2 and the importance sampler. We then analyze the runtime of Theorem 2, and the runtime and effective sample size (ESS) of the importance sampler.

4.3.1 Computing the action of a sparse matrix exponential

Both Theorem 2 and the importance sampler rely on "the action of the matrix exponential" (Al-Mohy and Higham, 2011). Let $\mathbf{A}$ be a $k \times k$ matrix and $\mathbf{v}$ a $1 \times k$ row vector. We need to compute expressions of the form $\mathbf{v}e^{\mathbf{A}}$. Naively, this kind of vector-matrix multiplication costs $O(k^2)$. However, in our case $\mathbf{A}$ will be sparse, with k nonzero entries, allowing us to more efficiently compute $\mathbf{v}e^{\mathbf{A}}$.

In particular, we use the algorithm of Al-Mohy and Higham (2011), as implemented in the Python package `scipy`. For $s \in \mathbb{Z}_+$, define $T_m(s^{-1}\mathbf{A}) = \sum_{i=0}^{m} (s^{-1}\mathbf{A})^i/i!$, the truncated Taylor series approximation of $e^{s^{-1}\mathbf{A}}$. Then, we have

$$\mathbf{v}e^{\mathbf{A}} = \mathbf{v}\left(e^{s^{-1}\mathbf{A}}\right)^s \approx \mathbf{v}[T_m(s^{-1}\mathbf{A})]^s.$$

Now let $\mathbf{b}_j = \mathbf{v}[T_m(s^{-1}\mathbf{A})]^j$, so B_j is a $1 \times k$ row vector. Then

$$\mathbf{b}_j = \mathbf{b}_{j-1}T_m(s^{-1}\mathbf{A}) = \sum_{i=0}^{m} \mathbf{b}_{j-1}\frac{(s^{-1}\mathbf{A})^i}{i!},$$

with $\mathbf{v}e^{\mathbf{A}} \approx \mathbf{b}_s$, and $\mathbf{b}_s$ evaluated in $\mathcal{T} = ms$ vector-matrix multiplications, each costing $O(k)$ by the sparsity of $\mathbf{A}$. Approximating $\mathbf{v}e^{\mathbf{A}}$ thus costs $O(\mathcal{T}k)$ time. Both m, s are chosen automatically to bound

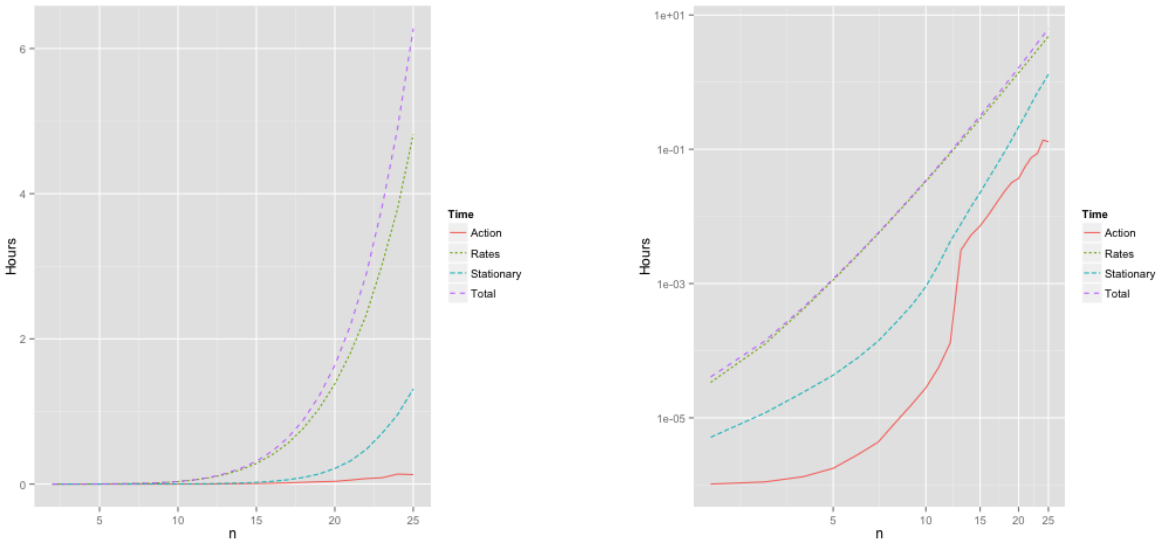
$$\frac{\|\Delta\mathbf{A}\|_1}{\|\mathbf{A}\|_1} \leq \text{tolerance} \approx 1.1 \times 10^{-16},$$

with $\Delta\mathbf{A}$ defined by $[T_m(s^{-1}\mathbf{A})]^s = e^{\mathbf{A}+\Delta\mathbf{A}}$. To avoid numerical instability, m is also bounded by $m \leq m_{\max} = 55$.

We note that $\mathbf{b}_j \approx \mathbf{v}e^{s^{-1}j\mathbf{A}}$, and thus this algorithm approximates $\mathbf{v}e^{t\mathbf{A}}$ along a grid of points $t \in \{s^{-1}, 2s^{-1}, \ldots, 1\}$. If $\mathbf{v}e^{t\mathbf{A}}$ is needed for additional $t \in J \subset (0, 1]$, then extra grid points can be added, to compute $\{\mathbf{v}e^{t\mathbf{A}}\}_{t\in J}$ in $\mathcal{T} + m|J|$ vector-matrix multiplications.

4.3.2 Runtime of the exact algorithm in Theorem 2

We consider the time complexity of computing $\mathbb{P}_0(\mathbf{n})$ via Theorem 2. Note that the formula (4.4) simultaneously computes $\mathbb{P}_0(\mathbf{n})$ for all configurations $\mathbf{n} \in \mathcal{N}$.



(a) (b)

Part	$\hat{\beta}$
Total	5.9
Construct rates $\tilde{\mathbf{\Lambda}}^d$	5.6
Initial stationary $\tilde{\boldsymbol{\lambda}}^{-D}$	8.1
Actions $(\mathbf{p}^d \odot \tilde{\boldsymbol{\gamma}}^d) e^{\tilde{\mathbf{\Lambda}}^d (t_{d+1} - t_d)}$	5.8

(c)

Figure 4.4: Runtime of Theorem 2 for $\theta = 0.008$, $\rho = 1.0$, and the demography described in (4.8). Includes total time, as well as times of subroutines, to construct the rate matrices $\tilde{\mathbf{\Lambda}}^d$, to compute the initial stationary $\tilde{\boldsymbol{\lambda}}^{-D}$, and to compute the actions $(\mathbf{p}^d \odot \tilde{\boldsymbol{\gamma}}^d) e^{\tilde{\mathbf{\Lambda}}^d (t_{d+1} - t_d)}$. Results were computed on a single core of a desktop computer (Mac Pro, c. early 2008) containing 16 GB of RAM. (a) Runtime for $n = 2, \ldots, 25$. For these n, constructing $\tilde{\mathbf{\Lambda}}^d$ dominates the runtime, and thus the total runtime looks like $O(n^6)$. (b) Same plot in log-log scale. A linear slope β in log-log scale corresponds to a polynomial degree β in original scale ($\log t = \beta \log n + \delta \Leftrightarrow t = Cn^\beta$). (c) Fit slope $\hat{\beta}$ to the last 10 points $n = 16, \ldots, 25$ of Figure 4.4b, via simple linear regression. Runtime is thus approximately $Cn^{\hat{\beta}}$.

For our analysis, we let $\mathcal{A} = \{0, 1\}$, as is assumed by LDhat and the applications in Section 4.4. We start by considering the dimensions of the vectors $\mathbf{p}^d$ and matrices $\tilde{\mathbf{\Lambda}}^d$ for intervals $(t_d, t_{d+1}]$. The set of a, b, c haplotypes is $H = \{00, 01, 10, 11, 0*, 1*, *0, *1\}$, so $|H| = 8$. Thus, there are $O(n^6)$ possible configurations $\mathbf{n}$ with $n^{(a)} = n^{(b)} = n - n^{(c)}$. In particular, there are $O(n^6)$ ways to specify $n_{00}, n_{01}, n_{10}, n_{11}, n_{0*}, n_{*0}$, and then $n_{1*} = n - \sum_{i,j\in\{0,1\}} n_{ij} - n_{0*}$ and $n_{*1} = n - \sum_{i,j\in\{0,1\}} n_{ij} - n_{*0}$ are determined. Thus, $\mathbf{p}^d$ is a row vector of dimension $1 \times O(n^6)$, and $\tilde{\mathbf{\Lambda}}^d$ is a square matrix of dimension $O(n^6) \times O(n^6)$, but $\tilde{\mathbf{\Lambda}}^d$ is sparse, with only $O(n^6)$ nonzero entries.

By using the algorithm of Section 4.3.1, we can compute $\mathbf{p}^{d+1} = \left[(\mathbf{p}^d \odot \tilde{\boldsymbol{\gamma}}^d) e^{\tilde{\mathbf{\Lambda}}^d (t_{d+1} - t_d)}\right] \div \tilde{\boldsymbol{\gamma}}^d$ from $\mathbf{p}^d$ in $O(\mathcal{T}_d n^6)$ time, where $\mathcal{T}_d$ is the number of vector-matrix multiplications to compute the action of $e^{\tilde{\mathbf{\Lambda}}^d (t_{d+1} - t_d)}$. We note that the stationary distribution $(\gamma_0^d, \ldots, \gamma_n^d)$ can be computed in $n+1$ steps: $\mathbf{\Gamma}^d$ is the rate matrix of a simple random walk with $n+1$ states, so $\gamma_{i+1}^d = \gamma_i^d \frac{[\Gamma^d]_{i,i+1}}{[\Gamma^d]_{i+1,i}}$ and $\sum_i \gamma_i^d = 1$.

Similarly, the initial value $\mathbf{p}^{-D+1} = \tilde{\boldsymbol{\lambda}}^{-D} \div \tilde{\boldsymbol{\gamma}}^{-D}$ can be computed via sparse vector-matrix multiplications, using the technique of power iteration. For $\mu = \frac{1}{\max_{ij} [\tilde{\mathbf{\Lambda}}^{-D}]_{ij}}$ and arbitrary positive vector $\mathbf{v}^{(0)}$ with $\|\mathbf{v}^{(0)}\|_1 = 1$, we have $\mathbf{v}^{(i)} := \mathbf{v}^{(0)} (\mu \tilde{\mathbf{\Lambda}}^{-D} + I)^i \to \tilde{\boldsymbol{\lambda}}^{-D}$ as $i \to \infty$. In particular, we set the number of iterations, $\mathcal{T}_{-D}$, so that $\left\| \log \mathbf{v}^{(\mathcal{T}_{-D})} \div \log \mathbf{v}^{(\mathcal{T}_{-D}-1)} \right\|_1 < 1 \times 10^{-8}$, where $\log \mathbf{v}^{(i)}$ is the element-wise log of $\mathbf{v}^{(i)}$.

To summarize, computing $\mathbb{P}_0(\mathbf{n})$ for all $O(n^6)$ configurations $\mathbf{n} \in \mathcal{N}$ of size n costs $O(n^6 \mathcal{T}_{\max} D)$, with $\mathcal{T}_{\max} = \max\{\mathcal{T}_{-D}, \ldots, \mathcal{T}_{-1}\}$. We caution that $\mathcal{T}_{\max}$ depends on $n, \{t_d\}, \{\tilde{\mathbf{\Lambda}}^d\}$.

Figure 4.4 shows the empirical runtime of Theorem 2 as a function of n. We used an example demography with $D = 3$ epochs, consisting of a sharp population bottleneck followed by a rapid expansion. Specifically, the population size history $\frac{1}{\alpha(t)}$, in coalescent-scaled units, is given by

$$\frac{1}{\alpha(t)} = \begin{cases} 100, & -0.5 < t \le 0, \\ 0.1, & -0.58 < t \le -0.5, \\ 1, & t \le -0.58. \end{cases} \tag{4.8}$$

We used $\theta = 0.008$, $\rho = 1.0$, and $n = 2, \ldots, 25$. For this range of n values, the runtime appears to be $\sim n^6$. However, the most expensive part of the computation was simply constructing the rate matrices $\{\tilde{\mathbf{\Lambda}}^d\}$, which we did with unoptimized Python code. By contrast, we used external modules `numpy` and `scipy` to compute all matrix operations, and these are heavily optimized.

Figure 4.4 was produced on a desktop computer (Mac Pro, c. early 2008) with 16 GB of RAM, and we did not exceed any memory limitations for the values of n we tried ($n \le 25$). The memory cost of Theorem 2 is $O(n^6)$, since $\tilde{\mathbf{\Lambda}}^d$ has $O(n^6)$ nonzero entries. In Section 4.4.2, we compute a lookup table for the same demography, with sample size $n = 20$ and 200 values of ρ. This took approximately 9 hours on a server using 20 cores, or about 1.1 hours per ρ on a single core.

4.3.3 Performance of importance sampler

Here we study the performance of our importance sampling scheme. We start by examining the computational complexity per importance sample. We then examine the empirical runtime and effective sample size for the example demography described in (4.8).

To construct our proposal distribution $\hat{Q}$, we must first precompute the Moran likelihoods $\mathbb{P}_{(N)}(\mathbf{M}_t)$ using (4.2) along a grid of points $t \in \{\tau_1, \tau_2, \ldots, \tau_J\}$. Again assuming a model with $|\mathcal{A}| = 2$ as in LDhat, $\mathbf{M}_t$ can take on $O(N^3)$ possible states. The transition rate matrix $\mathbf{\Lambda}_d^{(N)}$ thus has dimensions $O(N^3) \times O(N^3)$, but is sparse with $O(N^3)$ nonzero entries, and thus we can efficiently compute the action of its exponential, as discussed in Section 4.3.1, resulting in a cost of $O(N^3(\mathcal{T}D + J))$, with $\mathcal{T}$ being the number of vector-matrix multiplications needed per epoch.

We compute $\mathbb{P}_t^{(N)}(\mathbf{n})$ by subsampling from $\mathbb{P}_{(N)}(\mathbf{M}_t)$ as in (4.3), and thus set $N = 2n$, since $2n$ is the maximum number of individuals in $\mathbf{n}$ (because each of the original n lineages can recombine into two lineages). However, it is inefficient to use (4.3) directly to compute an approximation $\hat{\mathbb{P}}_t(\mathbf{n})$ for every value of $\mathbf{n}$. Instead, it is better to use the recursive formula $\hat{\mathbb{P}}_t(\mathbf{n}) = \sum_{\mathbf{m}} \hat{\mathbb{P}}_t(\mathbf{m})\mathbb{P}(\mathbf{n} \mid \mathbf{m})$, where the sum is over all configurations $\mathbf{m}$ obtained by adding an additional sample to $\mathbf{n}$.

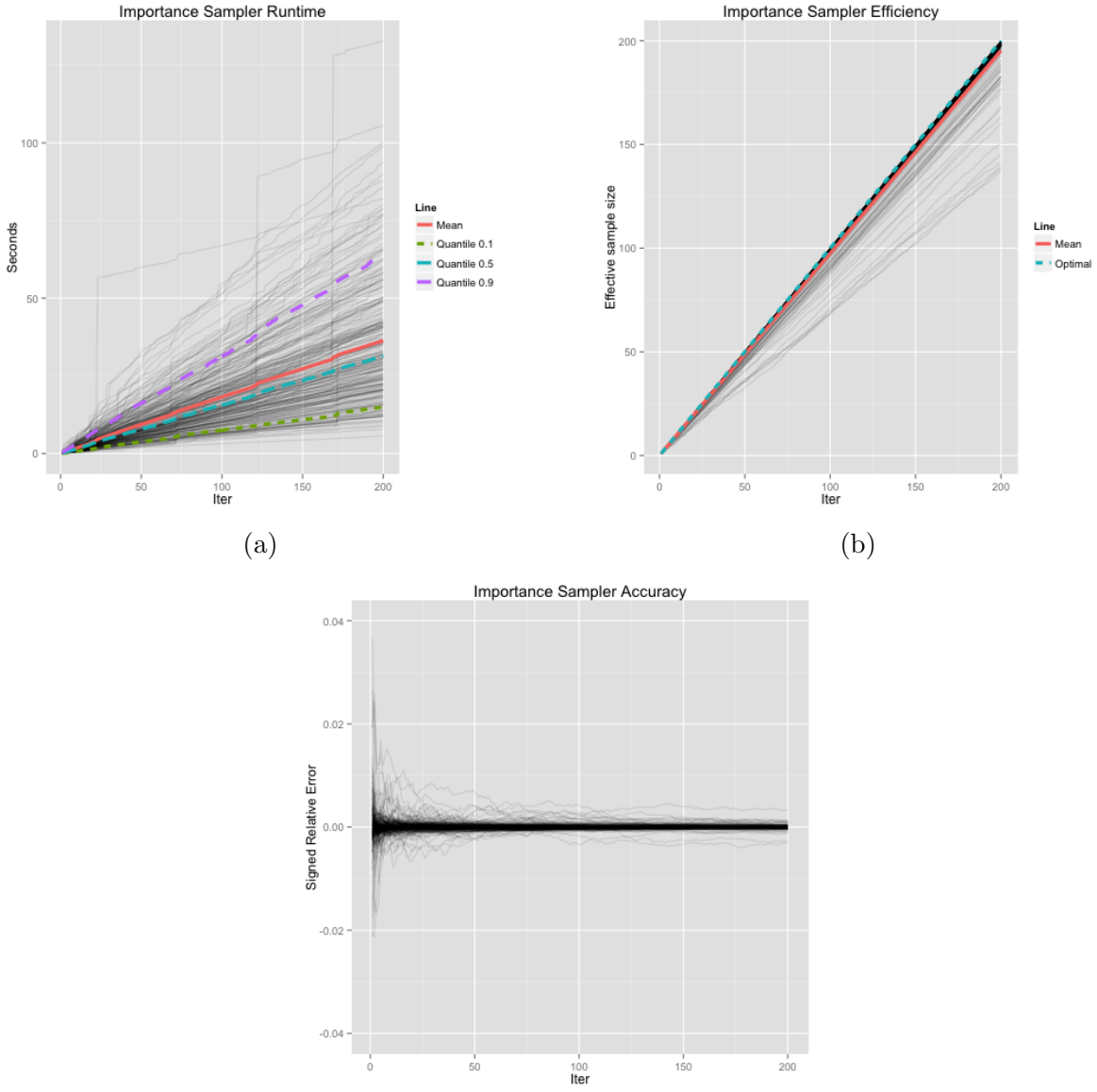
This costs $O(n^8 J)$ time and space, since there are J grid points and $O(n^8)$ possible configurations of $\mathbf{n}$. Then, assuming a reasonably efficient proposal, the expected cost to draw K importance samples is $O(nJK)$, since the expected number of coalescence, mutation, and recombination events before reaching the marginal common ancestor at each locus is $O(n)$ (Griffiths, 1991). This approach thus takes $O(n^3\mathcal{T}D + n^8 J + n^4 JK)$ expected time to compute $\mathbb{P}_0(\mathbf{n})$ for all $O(n^3)$ possible $\mathbf{n}$. In practice, we only precomputed $\hat{\mathbb{P}}_t(\mathbf{n})$ for the $O(n^4)$ fully specified $\mathbf{n}$ (without missing alleles), but computed and cached $\hat{\mathbb{P}}_t(\mathbf{n})$ as needed for partially specified $\mathbf{n}$ (with missing alleles). The theoretical running time to compute the full lookup table is still $O(n^3\mathcal{T}D+n^8 J+n^4 JK)$, but in practice, many values of $\mathbf{n}$ are highly unlikely and never encountered at each τ_j.

We examined the runtime of our importance sampler by computing $\mathbb{P}_0(\mathbf{n})$ for the same demographic history considered in the previous section, with $\theta = 0.008$ and $\rho = 1.0$, for all fully specified $\mathbf{n}$ with $n = 20$, drawing $K = 200$ genealogies per $\mathbf{n}$. Using a single processor on a server, computing $\mathbb{P}_t^{(N)}(\mathbf{M}_t)$ took 5.4 minutes with $J = 43$, while the additional precomputation took about 121.6 seconds. Sampling genealogies then took about 0.18 seconds per genealogy (Figure 4.5a).

The number K of importance samples required to reach a desired level of accuracy is typically measured with the effective sample size (ESS):

$$\text{ESS} = \frac{\left(\sum_{i=1}^K w_i\right)^2}{\sum_{i=1}^K w_i^2},$$

where w_i denotes the importance weight of the ith sample. Note that $\text{ESS} \leq K$ always, with equality only achieved if the w_i have 0 variance.



(a) (b)

(c)

Figure 4.5: Accuracy and runtime of importance sampling, on the demography described in (4.8) with $\theta = 0.008$ and $\rho = 1.0$, drawing 200 genealogies for each of the 275 fully-specified configurations $\mathbf{n}$ with $n = 20$. Results were run on a server using a single core and about 60 GB of RAM. (a) Runtime for each $\mathbf{n}$. The average time to sample 200 genealogies was 36.3 seconds, or 0.18 seconds per genealogy. (b) The ESS for each $\mathbf{n}$. The average ESS was 195.2 after 200 draws. (c) Relative error of $\log \hat{\mathbb{P}}(\mathbf{n})$, for each $\mathbf{n}$. The signed relative error is $\frac{\text{Est}-\text{Truth}}{\text{Truth}}$, with Truth computed via Theorem 2. The largest error ≈ 0.003 after 200 draws.

The previous two-locus importance sampler of Fearnhead and Donnelly (2001), which assumes a constant population size, achieves ESS anywhere between $0.05K$ and $0.5K$, depending on $\mathbf{n}, \theta, \rho$ (result not shown). This importance sampler is based on a similar result as Theorem 3, with optimal rates $\phi_{\mathbf{n},\mathbf{m}} \frac{\mathbb{P}(\mathbf{m})}{\mathbb{P}(\mathbf{n})}$. However, to approximate $\frac{\mathbb{P}(\mathbf{m})}{\mathbb{P}(\mathbf{n})}$, previous approaches did not use a Moran model, but followed the approach of Stephens and Donnelly (2000), using an approximate "conditional sampling distribution" (CSD). We initially tried using the CSD of Fearnhead and Donnelly (2001) and later generalizations to variable demography (Sheehan et al., 2013; Steinrücken et al., 2015), but found that importance sampling failed under population bottleneck scenarios, with the ESS repeatedly crashing to lower and lower values. Previous attempts to perform importance sampling under variable demography (Ye et al., 2013) have also encountered low ESS, though in the context of an infinite sites model without recombination.

By contrast, our importance sampler, with $\frac{\mathbb{P}_t(\mathbf{m})}{\mathbb{P}_t(\mathbf{n})}$ approximated via a Moran model, is extremely accurate: for the scenario in Figure 4.5, the average effective sample size (ESS) was about $0.976K$ (Figure 4.5b), and is very close to the true likelihood computed from Theorem 2 (Figure 4.5c), with the maximum relative error of $\log \hat{\mathbb{P}}(\mathbf{n})$ less than 0.3% after $K = 200$ samples.

4.4 Application

Previous simulation studies (McVean et al., 2002; Chan et al., 2012) have shown that if the demographic model is misspecified, composite-likelihood methods (which so far have assumed a constant population size) can produce recombination rate estimates that are biased. Many populations, including that of humans and *D. melanogaster*, have undergone bottlenecks in the recent past (Gutenkunst et al., 2009; Choudhary and Singh, 1987), and it has been argued (Johnston and Cutler, 2012) that such bottlenecks can severely affect recombination rate estimation, and can cause the appearance of spurious recombination hotspots. In this section, we examine to what extent correctly accounting for demography in the two-locus likelihoods improves fine-scale recombination rate estimation.

We first examine how a population bottleneck followed by rapid growth affects the correlation between partially linked sites. We then study how using the correct two-locus sampling probabilities affects recombination rate estimation under such a demographic model.

Throughout this section, we consider the example population size history $\frac{1}{\alpha(t)}$ described in (4.8). Under this model and $n = 2$, the expected time of common ancestor is $\mathbb{E}[T_{\text{MRCA}}] \approx 1$. We thus compare this demography against a constant size demography with coalescent-scaled size of $\frac{1}{\alpha} \equiv 1$, as this is the population size that would be estimated using the pairwise heterozygosity (Tajima, 1983).

In our example, we use a coalescent-scaled mutation rate of $\theta = 0.008$ per base, which is roughly the mutation rate of *D. melanogaster* (Chan et al., 2012). We use a mutation model with two alleles, labeled "0" and "1", which mutate to each other with the same rate θ.

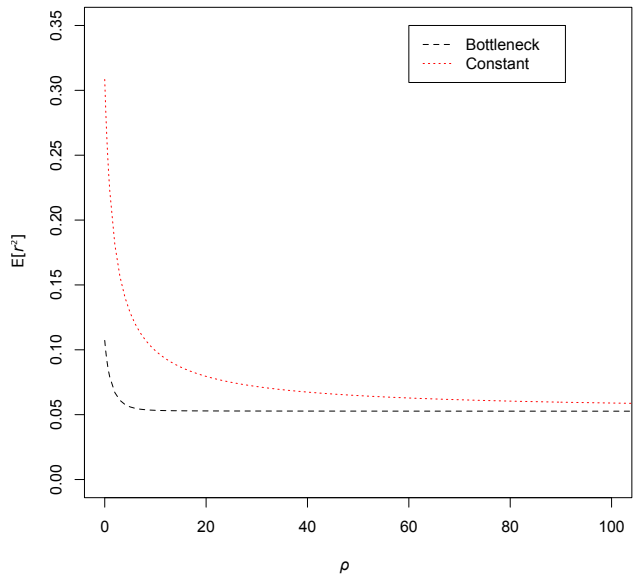


Figure 4.6: $\mathbb{E}[r^2]$ for the bottleneck and constant demographies, as a function of ρ. Under the bottleneck, even nearby sites are quite uncorrelated.

4.4.1 Linkage disequilibrium and two-locus likelihoods

Let $x_{ij} = \frac{n_{ij}}{n}$ be the fraction of haplotype ij in the sample, and let $x_i^{(a)} = \sum_j x_{ij}$ and $x_j^{(b)} = \sum_i x_{ij}$. A commonly used measure of linkage disequilibrium is

$$\mathbb{E}[r^2] = \mathbb{E}\left[\left(\frac{x_{11} - x_1^{(a)} x_1^{(b)}}{x_0^{(a)} x_1^{(a)} x_0^{(b)} x_1^{(b)}}\right)^2\right],$$

which corresponds to the expected square-correlation of a random allele at locus a with a random allele at locus b. The measure r^2 approximately follows a χ_1^2-distribution and can be used to test the statistical significance of linkage disequilibrium (Weir, 1996, p. 113). Figure 4.6 compares $\mathbb{E}[r^2]$ between the bottleneck model in (4.8) and the constant population size model. Under the bottleneck, $\mathbb{E}[r^2]$ is much lower for small ρ and decays more rapidly as $\rho \to \infty$. This implies that $\mathbb{E}[r^2]$ has less power to detect linkage under the bottleneck demography.

In Figure 4.7, we examine some specific $\mathbb{P}(\mathbf{n}; \rho)$ as a function of ρ. We note that for some configuration $\mathbf{n}$, the likelihood curves can have qualitatively different shapes under the bottleneck and constant demographies. To summarize the overall difference between the sampling distribution for the constant population size model and that for the bottleneck model, we show in Figure 4.8 the total variation (TV) distance between the two probability distributions conditioned on having both sites segregating. TV is bounded from above by 1, and so the sampling distributions are substantially different for all values of ρ.

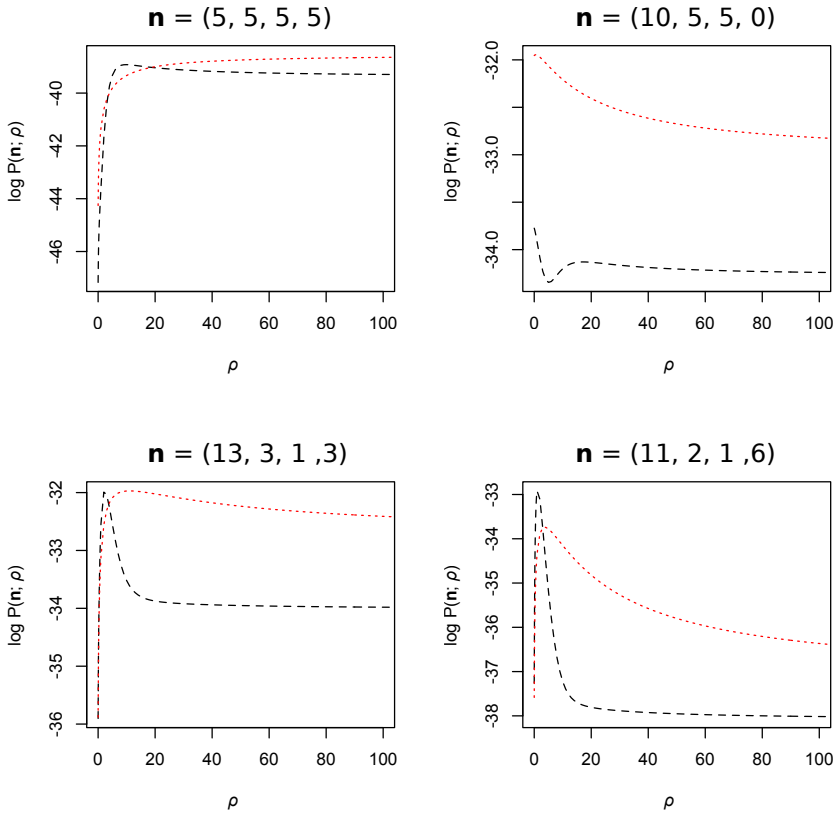


Figure 4.7: $\mathbb{P}(\mathbf{n};\rho)$ as a function of ρ, under the constant (red dotted) and bottleneck (black dashed) demographies, for some specific $\mathbf{n}$. Plots are labeled by $(n_{00}, n_{01}, n_{10}, n_{11})$. For some $\mathbf{n}$, $\log \mathbb{P}(\mathbf{n};\rho)$ has a qualitatively different shape under the constant and bottleneck demographies.

4.4.2 Fine-scale recombination rate estimation

For a sample of n haplotypes observed at L SNPs, let $\mathbf{n}[a,b]$ be the two-locus sample observed at SNPs $a, b \in \{1,\dots,L\}$, and let $\rho[a,b]$ be the recombination rate between SNPs a and b. Let W denote some window size, so that we only compute sampling probabilities at sites that are close enough, with $|b-a| < W$.

The programs LDhat (McVean et al., 2002, 2004; Auton and McVean, 2007), LDhot (Myers et al., 2005; Auton et al., 2014), and LDhelmet (Chan et al., 2012) use the composite likelihood

$$\hat{\mathcal{L}}(\boldsymbol{\rho}) = \prod_{a,b:b-a<W \text{ and } a<b} \mathbb{P}(\mathbf{n}[a,b];\rho[a,b]), \tag{4.9}$$

to estimate the fine-scale recombination map $\boldsymbol{\rho}$ and to infer recombination hotspots. However, they assume a constant population size history to compute $\mathbb{P}(\mathbf{n};\rho)$. Using simulations, Johnston and Cutler (2012) found that LDhat produces many spurious recombination

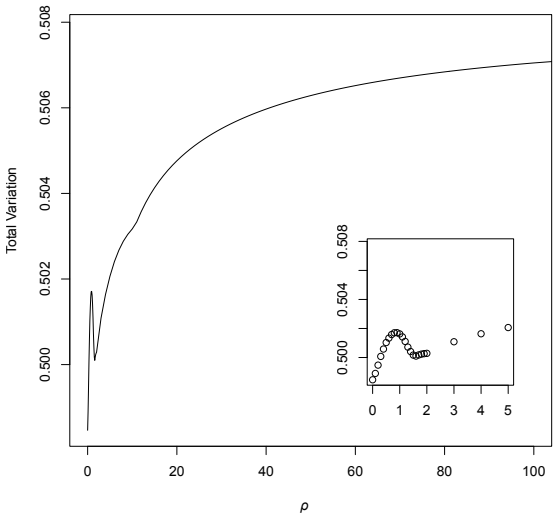


Figure 4.8: Total variation (TV) distance between the sampling distribution of the constant population size model and that of the bottleneck model conditioned on having both sites segregating, as a function of ρ. The inset shows the computed values of TV in more detail for smaller values of ρ.

hotspots when the true demographic history has an extreme bottleneck followed by rapid growth.

One problem with (4.9) is that $\hat{\boldsymbol{\mathcal{L}}}(\boldsymbol{\rho})$ is sharply peaked compared to the true likelihood, and thus prone to overfitting, as illustrated by the spurious hotspots found by Johnston and Cutler (2012). The problem of overfitting remains regardless of whether the correct size history $\frac{1}{\alpha(t)}$ is used. In order to get less noisy estimates of $\boldsymbol{\rho}$, we modified LDhelmet to flatten the composite likelihood, replacing (4.9) with

$$\hat{\boldsymbol{\mathcal{L}}}(\boldsymbol{\rho}) = \left[\prod_{a,b:b-a<W \text{ and } a<b} \mathbb{P}(\mathbf{n}[a,b]; \boldsymbol{\rho}[a,b]) \right]^{\frac{1}{W-1}}, \tag{4.10}$$

as done in Auton and McVean (2007). This corrects for the fact that each locus is contained in $W-1$ two-locus likelihoods. We found that using (4.10) in LDhelmet obviated the need for tuning the block penalty parameter in the program.

Using LDhelmet modified with (4.10), we investigated whether using the two-locus likelihood lookup table under the true demographic model improves the accuracy of the estimated map $\hat{\boldsymbol{\rho}}$. We divided the recombination map for the X chromosome of *Drosophila melanogaster* from Raleigh, NC inferred by Chan et al. (2012) into 22 non-overlapping 1Mb regions. For each of these regions we simulated 5 datasets with 20 individuals using MaCS (Chen et al., 2009). On each of these 110 datasets, we ran our modified version of LDhelmet using both the true bottleneck model and the misspecified constant demographic model. In Figure 4.9,

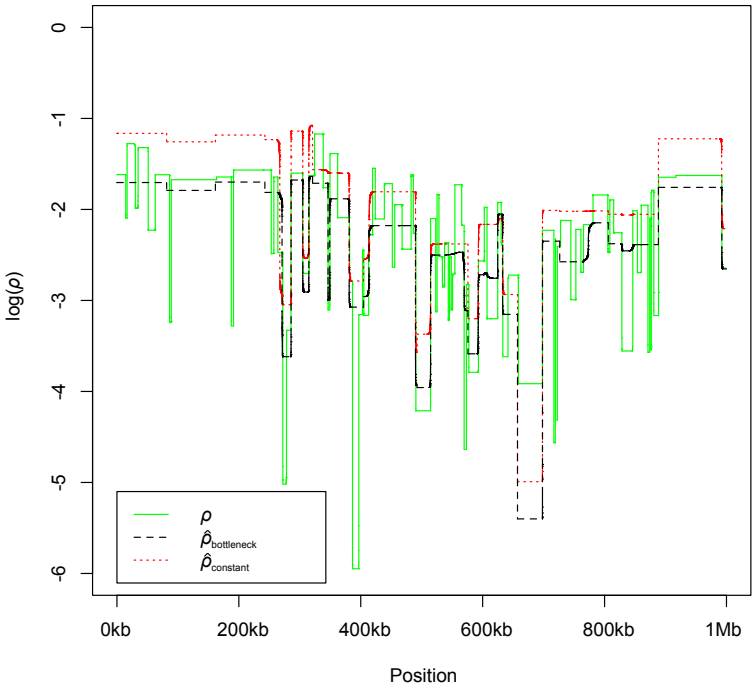


Figure 4.9: Posterior median $\hat{\boldsymbol{\rho}}$ inferred by LDhelmet, for a 1 Mb region simulated with MaCS (Chen et al., 2009) under the bottleneck demography described in (4.8) and using the recombination map (green) inferred by Chan et al. (2012) for a 1 Mb interior region of *D. melanogaster* X chromosome. The map $\hat{\boldsymbol{\rho}}_{\text{bottleneck}}$ (black dashed) is the inferred recombination map when the two-locus likelihood lookup table is computed assuming the true demographic model, while $\hat{\boldsymbol{\rho}}_{\text{constant}}$ (red dotted) is the map obtained when incorrectly assuming a constant population size history.

we compare the estimated map, $\hat{\boldsymbol{\rho}}$, obtained using the correct lookup table and that obtained using the lookup table of a misspecified model (constant population size). Table 4.4 and Figure 4.10 show the L_2 error, and the correlation of the true map, $\boldsymbol{\rho}$, and an estimated map, $\hat{\boldsymbol{\rho}}$, at different scales. While the performance of LDhelmet varied significantly depending on the underlying true recombination map, accounting for demography nearly always improved the per-base L_2 error, and improved the correlation at the fine scale and broader scales between the inferred and true maps in more than 70% of simulations. For some simulations, the improvements in L_2 error or correlation were dramatic. The improvement of all these statistics was significant according to a Wilcoxon signed-rank test ($p < 1 \times 10^{-6}$).

	$\frac{\|\|\hat{\boldsymbol{\rho}}-\boldsymbol{\rho}\|\|_2^2}{L}$	$\widehat{\text{Cor}}_{1\text{bp}}(\rho,\hat{\rho})$	$\widehat{\text{Cor}}_{1\text{kb}}(\rho,\hat{\rho})$	$\widehat{\text{Cor}}_{10\text{kb}}(\rho,\hat{\rho})$	$\widehat{\text{Cor}}_{30\text{kb}}(\rho,\hat{\rho})$
$\hat{\boldsymbol{\rho}}_{\text{constant}}$	0.000908	0.389	0.416	0.534	0.622
$\hat{\boldsymbol{\rho}}_{\text{bottleneck}}$	0.000686	0.441	0.472	0.597	0.679
p-value	2.2×10^{-16}	3.32×10^{-12}	2.37×10^{-12}	1.56×10^{-10}	5.75×10^{-7}

Table 4.4: Accuracy of estimated recombination map $\hat{\boldsymbol{\rho}}$, when assuming a constant population size and when accounting for the bottleneck. We simulated 110 datasets using 1Mb regions of the inferred recombination map of the X chromosome of *Drosophila melanogaster* computed by Chan et al. (2012). $\frac{||\rho-\hat{\rho}||_2^2}{L}$ is the per-base L_2 error. We also show the correlation of $\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}$ at different scales, as in Wegmann et al. (2011). All statistics were computed using only the middle 500Kb of each region to ameliorate edge effects. That is, $\widehat{\text{Cor}}_B(\rho,\hat{\rho})$ is the correlation of the true and estimated recombination rates over a physical distance of B bases, evaluated at the positions 250 Kb, 250 Kb $+ B$, 250 Kb $+ 2B, \ldots,$ 750 Kb. The statistics were computed for each dataset and then averaged over all 110 datasets. The p-values are for the null hypothesis that accounting for the bottleneck does not improve accuracy, under a Wilcoxon signed-rank test. The results are shown visually in Figure 4.10.

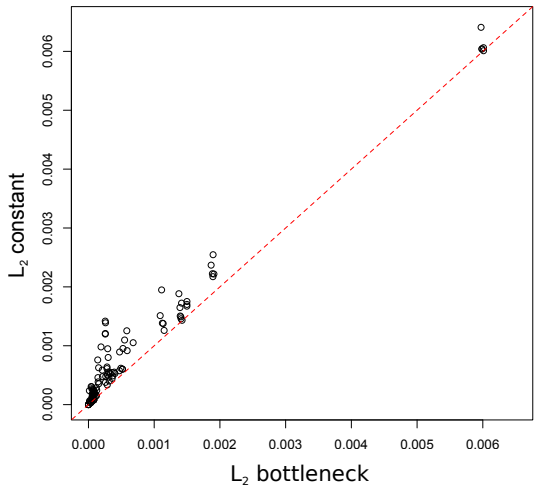
4.5 Proofs

For a stochastic process $\{X_t\}_{t\leq 0}$, we denote its partial sample paths with the following notation: $X_{s:t} = \{X_u : u \in (s,t]\}$ and $X_{\leq s} = X_{-\infty:s}$.

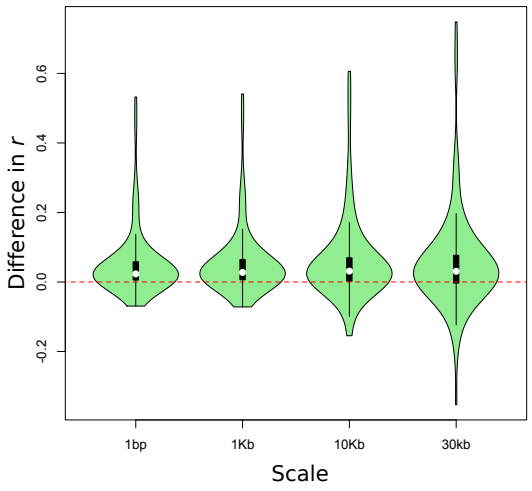
4.5.1 Proof of Theorem 2

Define C_t^* to be the number of c type lineages at time t, governed by the *backward-in-time* Markov chain with rate matrix $\boldsymbol{\Gamma}^d$ in the time interval $(t_d, t_{d+1}]$. For $s < t$, the partial sample path $\tilde{\mathbf{M}}_{s:t}^*$ of the augmented Moran process is generated by dropping down mutation and copying events (as described below) onto the sample path $C_{s:t}^*$ of coalescence and recombination events, and then evolving $\tilde{\mathbf{M}}_s^*$ forward in time (Figure 4.3). We illustrate the conditional independence structure of $\tilde{\mathbf{M}}_t^*$ and C_t^* via a directed graphical model (Koller and Friedman, 2009) in Figure 4.11. Mutation and copying events in $\tilde{\mathbf{M}}_{s:t}^*$ occur as follows:

- Mutations hit each individual at rate $\frac{\theta}{2}$ per locus. Alleles mutate according to transition matrix $\mathbf{P}$.
- The rate of copying events depends on the lineage type. Each pair of a types experiences a copying event at rate α_d, with the direction of copying chosen with probability $\frac{1}{2}$. The rates are the same for every pair of b and every pair of c types. Each pair of a and c types, as well as each pair of b and c types, also experience copying at rate α_d, but the direction of copying is always from the c type to the a or b type, and only happens at one allele (left for a, right for b). No pair of a and b types experiences copying events.



(a) The per-base L_2 error $\frac{\|\boldsymbol{\rho}-\hat{\boldsymbol{\rho}}\|^2}{L}$ using the true bottleneck demography vs. assuming a constant demography. Accounting for the bottleneck nearly always reduced the L_2 error, often substantially.



(b) The improvement in correlation $\widehat{\mathrm{Cor}}_B(\rho, \hat{\rho}_{\text{bottleneck}}) - \widehat{\mathrm{Cor}}_B(\rho, \hat{\rho}_{\text{constant}})$ from accounting for the bottleneck. The correlation improved in at least 70% of simulations at each scale B, occasionally dramatically.

Figure 4.10: Visual comparison of the statistics in Table 4.4, over each of the 110 simulations.

For $\mathbf{n}$ with $n^{(a)} = n^{(b)} = n - n^{(c)}$, the sampling probability $\mathbb{P}_t(\mathbf{n})$ defined in (4.1) for the coalescent with recombination satisfies

$$\mathbb{P}_t(\mathbf{n}) \quad = \quad \mathbb{P}(\tilde{\mathbf{M}}_t^* = \mathbf{n} \mid C_t^* = n^{(c)}). \tag{4.11}$$

To see why this is true, we trace the ancestry of $\tilde{\mathbf{M}}_t^*$ backwards in time (Figure 4.3b). When tracing past a Moran copying event, where a lineage x copies onto a lineage y, the ancestry of x and y coalesce: x becomes ancestral to all the haplotypes that traced their ancestry through y, and the Moran lineage y ceases to be ancestral to the sample. This backwards tracing induces an ARG on the lineages of $\tilde{\mathbf{M}}_t^*$. (4.11) then follows from observing that the coalescent with recombination gives the distribution of the induced ARG: moving backwards in time, coalescence/copying events are encountered at rate α_d per pair, and recombination events at rate $\frac{\rho}{2}$ per lineage. Furthermore, the allele of the MRCA follows the stationary distribution of $\mathbf{P}$, and alleles mutate at rate $\frac{\theta}{2}$ with transition matrix $\mathbf{P}$.

Now, define C_t to be the *foward-in-time* Markov chain with rate matrix $\mathbf{\Gamma}^d$ in $(t_d, t_{d+1}]$ (whereas C_t^* has the same rates but going *backward-in-time*). Let $\tilde{\mathbf{M}}_t$ be the forward-in-time Markov chain with conditional law

$$\mathbb{P}(\tilde{\mathbf{M}}_{\leq 0} \mid C_{\leq 0} = \mathcal{C}) = \mathbb{P}(\tilde{\mathbf{M}}_{\leq 0}^* \mid C_{\leq 0}^* = \mathcal{C}). \tag{4.12}$$

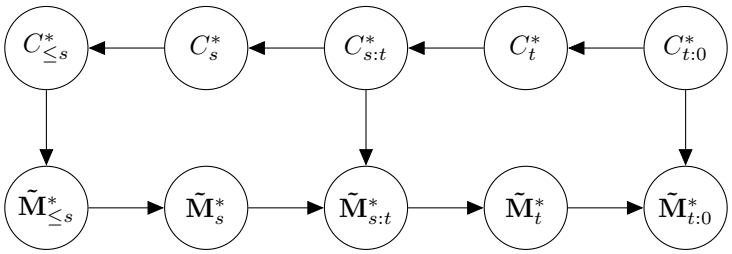


Figure 4.11: Probabilistic graphical model for the processes C^*_t and $\tilde{\mathbf{M}}^*_t$, with $-\infty < s < t \leq 0$.

Then, $\tilde{\mathbf{M}}_t$ has rate matrix $\tilde{\mathbf{\Lambda}}^d$ in $(t_d, t_{d+1}]$. The value of $\tilde{\mathbf{M}}_t$ at time t is a sample configuration that changes due to the mutation and copying events described above, and the following two additional events:

- Each c type individual experiences recombination at rate $\frac{\rho}{2}$, at which point it splits into an a and b type. Note this is similar to the coalescent with recombination, however here recombination happens *forward in time*, while in the coalescent with recombination it happens at rate $\frac{\rho}{2}$ going *backwards in time*.
- Each pair of a and b types coalesce into a single c type at rate α_d. Again, this is similar to the coalescent with recombination, but here the event happens *forward in time* rather than *backwards in time*.

Using (4.11) and (4.12), we next observe

$$\begin{aligned}
\mathbb{P}_{t_{d+1}}(\mathbf{n}) &= \mathbb{P}(\tilde{\mathbf{M}}^*_{t_{d+1}} = \mathbf{n} \mid C^*_{t_{d+1}} = n^{(c)}) \\
&= \sum_{\mathbf{m}} \mathbb{P}(\tilde{\mathbf{M}}^*_{t_d} = \mathbf{m} \mid C^*_{t_d} = m^{(c)}) \mathbb{P}(C^*_{t_d} = m^{(c)} \mid C^*_{t_{d+1}} = n^{(c)}) \\
&\qquad \times \quad \mathbb{P}(\tilde{\mathbf{M}}^*_{t_{d+1}} = \mathbf{n} \mid C^*_{t_{d+1}} = n^{(c)}, C^*_{t_d} = m^{(c)}, \tilde{\mathbf{M}}^*_{t_d} = \mathbf{m}) \\
&= \sum_{\mathbf{m}} \mathbb{P}_{t_d}(\mathbf{m}) \mathbb{P}(C^*_{t_d} = m^{(c)} \mid C^*_{t_{d+1}} = n^{(c)}) \\
&\qquad \times \quad \mathbb{P}(\tilde{\mathbf{M}}_{t_{d+1}} = \mathbf{n} \mid C_{t_{d+1}} = n^{(c)}, C_{t_d} = m^{(c)}, \tilde{\mathbf{M}}_{t_d} = \mathbf{m}). \qquad (4.13)
\end{aligned}$$

Note that in the second equality, we used the conditional independence of $\tilde{\mathbf{M}}^*_{t_d}$ and $C^*_{t_{d+1}}$ given $C^*_{t_d}$, which follows from the graphical model of Figure 4.11 by setting $s = t_d$ and $t = t_{d+1}$.

Next, note that $\mathbf{\Gamma}^d$ is the transition matrix of a simple random walk with bounded state

space and no absorbing states, and thus is reversible. Thus,

$$\begin{aligned}\gamma^d_{n^{(c)}}\mathbb{P}(C^*_{t_d} = m^{(c)} \mid C^*_{t_{d+1}} = n^{(c)}) &= \gamma^d_{n^{(c)}}\left(e^{\Gamma^d(t_{d+1}-t_d)}\right)_{n^{(c)},m^{(c)}} \\ &= \gamma^d_{m^{(c)}}\left(e^{\Gamma^d(t_{d+1}-t_d)}\right)_{m^{(c)},n^{(c)}} \\ &= \gamma^d_{m^{(c)}}\mathbb{P}(C_{t_{d+1}} = n^{(c)} \mid C_{t_d} = m^{(c)}). \qquad (4.14)\end{aligned}$$

Plugging (4.14) into (4.13) yields

$$\begin{aligned}\mathbb{P}_{t_{d+1}}(\mathbf{n}) &= \sum_{\mathbf{m}} \mathbb{P}_{t_d}(\mathbf{m})\frac{\tilde{\gamma}^d_{\mathbf{m}}}{\tilde{\gamma}^d_{\mathbf{n}}}\mathbb{P}(C_{t_{d+1}} = n^{(c)} \mid C_{t_d} = m^{(c)}) \\ &\qquad \times \quad \mathbb{P}(\tilde{\mathbf{M}}_{t_{d+1}} = \mathbf{n} \mid C_{t_{d+1}} = n^{(c)}, C_{t_d} = m^{(c)}, \tilde{\mathbf{M}}_{t_d} = \mathbf{m}) \\ &= \sum_{\mathbf{m}} \mathbb{P}_{t_d}(\mathbf{m})\frac{\tilde{\gamma}^d_{\mathbf{m}}}{\tilde{\gamma}^d_{\mathbf{n}}}\mathbb{P}(C_{t_{d+1}} = n^{(c)}, \tilde{\mathbf{M}}_{t_{d+1}} = \mathbf{n} \mid C_{t_d} = m^{(c)}, \tilde{\mathbf{M}}_{t_d} = \mathbf{m}) \\ &= \sum_{\mathbf{m}} \mathbb{P}_{t_d}(\mathbf{m})\frac{\tilde{\gamma}^d_{\mathbf{m}}}{\tilde{\gamma}^d_{\mathbf{n}}}\left(e^{\tilde{\mathbf{\Lambda}}^d(t_{d+1}-t_d)}\right)_{\mathbf{m},\mathbf{n}}.\end{aligned}$$

which proves half of the desired result, i.e., $\mathbf{p}^{d+1} = \left((\mathbf{p}^d \odot \tilde{\boldsymbol{\gamma}}^d)e^{\tilde{\mathbf{\Lambda}}^d(t_{d+1}-t_d)}\right) \div \tilde{\boldsymbol{\gamma}}^d$, where $\mathbf{p}^d = [\mathbb{P}_{t_d}(\mathbf{n})]'_{\mathbf{n}}$. To show the other half, that $\mathbf{p}^{-D+1} = \tilde{\boldsymbol{\lambda}}^{-D} \div \tilde{\boldsymbol{\gamma}}^{-D}$, we simply note that for all $t \leq t_{-D+1}$,

$$\begin{aligned}\mathbb{P}_t(\mathbf{n})\tilde{\gamma}^{-D}_{\mathbf{n}} &= \mathbb{P}(\tilde{\mathbf{M}}^*_t = \mathbf{n} \mid C^*_t = n^{(c)})\gamma^{-D}_{n^{(c)}} \\ &= \mathbb{P}(\tilde{\mathbf{M}}_t = \mathbf{n} \mid C_t = n^{(c)})\gamma^{-D}_{n^{(c)}} \\ &= \mathbb{P}(\tilde{\mathbf{M}}_t = \mathbf{n}) \\ &= \tilde{\lambda}^{-D}_{\mathbf{n}},\end{aligned}$$

where the second equality follows by reversibility of $\mathbf{\Gamma}^{-D}$, which implies $\mathbb{P}(C_{\leq t} \mid C_t) = \mathbb{P}(C^*_{\leq t} \mid C^*_t)$, and thus $\mathbb{P}(\tilde{\mathbf{M}}_{\leq t} \mid C_t) = \mathbb{P}(\tilde{\mathbf{M}}^*_{\leq t} \mid C^*_t)$.

4.5.2 Proof of Theorem 3

We first check that $\mathbb{P}(\mathbf{n}_{s_1} \mid \mathbf{n}_{s_2}, \mathbf{n}_{s_3}) = \mathbb{P}(\mathbf{n}_{s_1} \mid \mathbf{n}_{s_2})$, for $-\infty < s_1 < s_2 < s_3 \leq 0$, and so $\mathbf{n}_t$ is a backwards in time Markov chain.

Recall that we generate $n^{(abc)}_t$ as a backwards in time Markov chain, then generate $\mathbf{n}_t$ by dropping down mutations forward in time. The conditional independence structure of $\mathbf{n}_{s_1}, \mathbf{n}_{s_2}, \mathbf{n}_{s_3}$ is thus described by the directed graphical model (Koller and Friedman, 2009) in Figure 4.12.

Doing moralization and variable elimination (Koller and Friedman, 2009) on Figure 4.12 results in the undirected graphical model in Figure 4.13. The graphical model of Figure 4.13

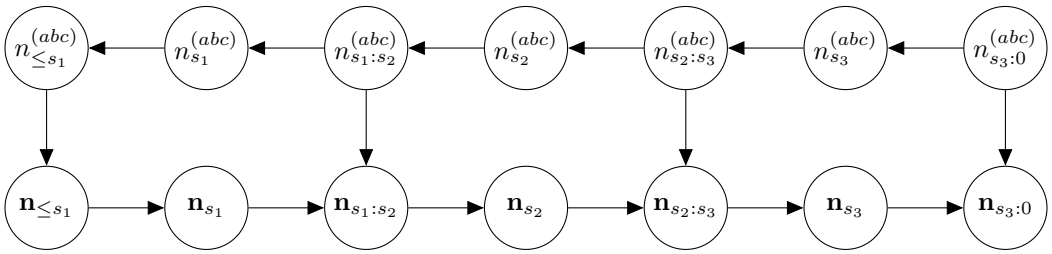


Figure 4.12: Probabilistic graphical model for the coalescent with recombination and mutation, with $-\infty < s_1 < s_2 < s_3 \leq 0$.

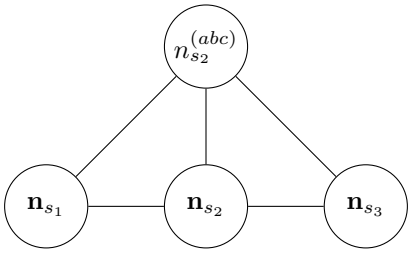


Figure 4.13: Undirected graphical model, after moralization and variable elimination on Figure 4.12: we add edges to form cliques on the left and right sides of $n^{(abc)}_{s_2}, \mathbf{n}_{s_2}$, and then eliminate all the variables except the ones pictured here.

then implies

$$\begin{aligned}\mathbb{P}(\mathbf{n}_{s_1} \mid \mathbf{n}_{s_2}, \mathbf{n}_{s_3}) &= \sum_{n^{(abc)}_{s_2}} \mathbb{P}(\mathbf{n}_{s_1} \mid \mathbf{n}_{s_2}, n^{(abc)}_{s_2}) \mathbb{P}(n^{(abc)}_{s_2} \mid \mathbf{n}_{s_2}, \mathbf{n}_{s_3}) \\ &= \mathbb{P}(\mathbf{n}_{s_1} \mid \mathbf{n}_{s_2}),\end{aligned}$$

where the second equality follows because $n^{(abc)}_{s_2}$ is a deterministic function of $\mathbf{n}_{s_2}$. Thus, $\mathbf{n}_t$ is a backwards in time Markov chain.

We next compute the backwards in time rates $q^{(t)}_{\mathbf{n},\mathbf{m}}$ for the Markov chain $\mathbf{n}_t$ at time t.

Starting from the definition of $q^{(t)}_{\mathbf{n},\mathbf{m}}$,

$$\begin{aligned}
q^{(t)}_{\mathbf{n},\mathbf{m}} &= \frac{d}{ds}\mathbb{P}(\mathbf{n}_{t-s} = \mathbf{m} \mid \mathbf{n}_t = \mathbf{n})\Big|_{s=0} \\
&= \frac{d}{ds}\frac{\mathbb{P}(\mathbf{n}_{t-s} = \mathbf{m}, \mathbf{n}_t = \mathbf{n} \mid n^{(abc)}_t = n^{(abc)})}{\mathbb{P}(\mathbf{n}_t = \mathbf{n} \mid n^{(abc)}_t = n^{(abc)})}\Big|_{s=0} \\
&= \frac{1}{\mathbb{P}_t(\mathbf{n})}\frac{d}{ds}\Big[\mathbb{P}(n^{(abc)}_{t-s} = m^{(abc)} \mid n^{(abc)}_t = n^{(abc)}) \\
&\qquad \times \mathbb{P}(\mathbf{n}_t = \mathbf{n} \mid n^{(abc)}_t = n^{(abc)}, \mathbf{n}_{t-s} = \mathbf{m})\mathbb{P}_{t-s}(\mathbf{m})\Big]\Big|_{s=0} \\
&= \frac{1}{\mathbb{P}_t(\mathbf{n})}\Big[\mathbb{P}(n^{(abc)}_t = m^{(abc)} \mid n^{(abc)}_t = n^{(abc)})\mathbb{P}(\mathbf{n}_t = \mathbf{n} \mid n^{(abc)}_t = n^{(abc)}, \mathbf{n}_t = \mathbf{m}) \\
&\qquad \times \frac{d}{ds}\mathbb{P}_{t-s}(\mathbf{m})\Big|_{s=0} + \phi^{(t)}_{\mathbf{n},\mathbf{m}}\mathbb{P}_t(\mathbf{m})\Big] \\
&= \begin{cases} \phi^{(t)}_{\mathbf{n},\mathbf{m}}\frac{\mathbb{P}_t(\mathbf{m})}{\mathbb{P}_t(\mathbf{n})}, & \text{if } \mathbf{m} \neq \mathbf{n}, \\ \phi^{(t)}_{\mathbf{n},\mathbf{n}} - \frac{d}{dt}\log \mathbb{P}_t(\mathbf{n}), & \text{if } \mathbf{m} = \mathbf{n}, \end{cases}
\end{aligned}$$

where the penultimate equality follows from the product rule and the definition of $\boldsymbol{\phi}^{(t)}$ in (4.6).

The specific entries of $\boldsymbol{\phi}^{(t)}$ listed in Table 4.3 can be obtained by applying the product rule to (4.6), and noting that $\frac{d}{ds}\mathbb{P}(n^{(abc)}_{t-s} \mid n^{(abc)}_t)|_{s=0}$ and $\frac{d}{ds}\mathbb{P}(\mathbf{n}_t \mid n^{(abc)}_t, \mathbf{n}_{t-s})|_{s=0}$ are, respectively, the backwards in time rates of $n^{(abc)}_t$ (as listed in Table 4.1), and the forward in time rates for dropping mutations on $\mathbf{n}_t$.

Chapter 5

Future directions

Computing the coalescent sampling probabilities at one or two loci, for a fully general model of demography and selection, remains a challenging problem. In this dissertation, we made mathematical and computational progress within the context of a variable demographic history and a neutral model of selection. We now discuss some remaining open problems and future directions.

- Can we add natural selection to our methods? The Moran model can be naturally extended to include purifying selection (Durrett, 2008; Donnelly and Kurtz, 1999), but this model is not exactly equal to the coalescent, except in the limit of an infinite number of lineages. Thus, to include purifying selection, we must find new, finite representations of the Moran model that exactly model the coalescent with selection. Alternatively, we could simply use the existing Moran model with selection, but with enough lineages so that the model is a good approximation to the coalescent.

- The effect of positive selection, in particular genetic "hitchhiking", can be modeled by Λ- and Ξ-coalescents, which are coalescent models with multiple simultanenous mergers. Λ- and Ξ- coalescents also model the effects of large family sizes, where a single individual is the parent of a large fraction of the population (for example, this occurs in some marine species). It should be relatively straightforward to extend our method for the multipopulation SFS to include Λ- and Ξ-coalescents, especially in light of recent results by Spence, Kamm, and Song (2015).

- In Chapter 4, we computed the two locus sampling probability for a single population with changing size. It would be interesting to extend this to multiple populations, by combining these results with the methods of Chapters 2 and 3. However, the computational complexity may become too large for practical application. Instead, it would be much faster to use a multipopulation version of the standard 2-locus Moran model, without augmenting it or using importance sampling to correct for its deviation from the coalescent. The results of Chapter 4, in particular the high ESS, suggest that

this should be reasonably accurate; furthermore, this approach would converge to the correct coalescent likelihood as the number of lineages goes to infinity.

- So far, computing the SFS under continuous migration is only possible under a diffusion (Gutenkunst et al., 2009) or Monte Carlo (Excoffier et al., 2013) approach. Can we extend our dynamic program, in particular the results of Chapter 3, to include continuous migration? It may be possible to do this by taking a continuous limit of our formula for discrete admixture events.

- A crucial issue is *identifiability*, or when it is possible to distinguish between alternative population histories. Specifically, if the sampling probability is exactly the same for two demographic histories, then we cannot distinguish between the two, and the problem of demographic inference is not identifiable. Identifiability conditions for the SFS have been established for the size history of a single population (Bhaskar and Song, 2014), but are lacking for multiple populations, and for the sampling probability at two loci.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

Awad H. Al-Mohy and Nicholas J. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM Journal on Scientific Computing*, 33 (2):488–511, 2011.

David J Aldous. Exchangeability and related topics. In P.L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin Heidelberg, 1985.

Adam Auton and Gil McVean. Recombination rate estimation in the presence of hotspots. *Genome research*, 17(8):1219–1227, 2007.

Adam Auton, Adi Fledel-Alon, Susanne Pfeifer, Oliver Venn, Laure Ségurel, Teresa Street, Ellen M Leffler, Rory Bowden, Ivy Aneas, John Broxholme, et al. A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336(6078):193–198, 2012.

Adam Auton, Ying Rui Li, Jeffrey Kidd, Kyle Oliveira, Julie Nadel, J Kim Holloway, Jessica J Hayward, Paula E Cohen, John M Greally, Jun Wang, et al. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genetics*, 9(12):e1003984, 2013.

Adam Auton, Simon Myers, and Gil McVean. Identifying recombination hotspots using population genetic data. March 2014.

F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327:836–840, 2010.

I. L. Berg, R. Neumann, K. G. Lam, S. Sarbajna, L. Odenthal-Hesse, C. A. May, and A. J. Jeffreys. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, 42(10):859–863, 2010.

A Bhaskar and Y. S. Song. Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability*, 44:391–407, 2012. (PMC3409093).

A Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics*, 42(6):2469–2493, 2014.

A. Bhaskar, J. A. Kamm, and Y. S. Song. Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability*, 44:408–428, 2012. (PMC3953561).

A. Bhaskar, Y. X. Rachel Wang, and Y. S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2):268–279, 2015.

David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A. Rosenberg, and Arindam RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932, 2012.

C Cannings. The latent roots of certain markov chains arising in genetics: a new approach, I. haploid models. *Advances in Applied Probability*, 6:260–290, 1974.

Andrew H Chan, Paul A. Jenkins, and Yun S. Song. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. *PLoS Genetics*, 8(12):e1003090, December 2012.

Gary K. Chen, Paul Marjoram, and Jeffrey D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res.*, 19:136–142, 2009.

Hua Chen. The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology*, 81(2):179–195, 2012.

Hua Chen. Intercoalescence time distribution of incomplete gene genealogies in temporally varying populations, and applications in population genetic inference. *Annals of Human Genetics*, 77(2):158–173, 2013.

M Choudhary and RS Singh. Historical effective size and the level of genetic diversity in drosophila melanogaster and drosophila pseudoobscura. *Biochemical genetics*, 25(1-2): 41–51, 1987.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Alex Coventry, Lara M Bull-Otterson, Xiaoming Liu, Andrew G Clark, Taylor J Maxwell, Jacy Crosby, James E Hixson, Thomas J Rea, Donna M Muzny, Lora R Lewis, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1:131, 2010.

M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories. I. *Adv. Appl. Prob.*, 36:417–433, 2004.

Nicola De Maio, Christian Schlötterer, and Carolin Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular biology and evolution*, 30(10):2249–2262, 2013.

Peter Donnelly and TG Kurtz. A countable representation of the Fleming-Viot measure-valued diffusion. *Ann Probab*, 24:698–742, 1996.

Peter Donnelly and Thomas G Kurtz. Genealogical processes for fleming-viot models with selection and recombination. *Annals of Applied Probability*, pages 1091–1148, 1999.

Peter Donnelly, Thomas G Kurtz, et al. Particle representations for measure-valued population models. *The Annals of Probability*, 27(1):166–205, 1999.

R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York, 2nd edition, 2008.

S. N. Ethier and R. C. Griffiths. On the two-locus sampling distribution. *J. Math. Biol.*, 29: 131–159, 1990.

Stewart N Ethier and Thomas G Kurtz. Fleming-viot processes in population genetics. *SIAM Journal on Control and Optimization*, 31(2):345–386, 1993.

Warren J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

Warren J. Ewens. *Mathematical Population Genetics: I. Theoretical Introduction*. Springer Science+Business Media, Inc., New York, 2004.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10): e1003905, 2013.

J C Fay and C I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155:1405–1413, 2000.

P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

P. Fearnhead and N. G. C. Smith. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.*, 77:781–794, 2005.

P. Fearnhead, R. M Harding, J. A. Schneider, S. Myers, and P. Donnelly. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, 167:2067–2081, 2004.

Paul Fearnhead. SequenceLDhot: detecting recombination hotspots. *Bioinformatics*, 22: 3061–3066, 2006.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

Elodie Gazave, Li Ma, Diana Chang, Alex Coventry, Feng Gao, Donna Muzny, Eric Boerwinkle, Richard A Gibbs, Charles F Sing, Andrew G Clark, et al. Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences*, 111(2):757–762, 2014.

G. B. Golding. The sampling distribution of linkage disequilibrium. *Genetics*, 108:257–274, 1984.

Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, David L Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

Andreas Griewank and George F Corliss. *Automatic differentiation of algorithms: theory, implementation, and application*. Society for industrial and Applied Mathematics Philadelphia, PA, 1991.

R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in population genetics and human evolution*, volume 87, pages 257–270. Springer-Verlag, Berlin, 1997.

R. C. Griffiths, P. A. Jenkins, and Y. S. Song. Importance sampling and the two-locus model with subdivided population structure. *Advances in Applied Probability*, 40:473–500, 2008.

R.C. Griffiths. The two-locus ancestral graph. *Selected Proceedings of the Sheffield Symposium on Applied Probability. IMS Lecture Notes–Monograph Series*, 18:100–117, 1991.

R.C. Griffiths and Simon Tavaré. The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, 14(1-2):273–295, 1998.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, 2009.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM: Society for Industrial and Applied Mathematics, 2nd edition, 2002.

Kent E Holsinger and Bruce S Weir. Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10(9):639–650, 2009.

F. Hoppe. Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, 20:91–94, 1984.

R. R. Hudson. Generating samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

R.R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4): 1805–1817, 2001.

P. A. Jenkins and Y. S. Song. Closed-form two-locus sampling distributions: accuracy and universality. *Genetics*, 183:1087–1103, 2009.

P. A. Jenkins and Y. S. Song. An asymptotic sampling formula for the coalescent with recombination. *Annals of Applied Probability*, 20:1005–1028, 2010.

Paul A. Jenkins and Yun S. Song. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology,*, 80:158–173, 2011.

Paul A. Jenkins and Yun S. Song. Padé approximants and exact two-locus sampling distributions. *Annals of Applied Probability*, 22:576–607, 2012.

Paul A Jenkins, Jonas W Mueller, and Yun S Song. General triallelic frequency spectrum under demographic models with variable population size. *Genetics*, 196(1):295–311, 2014.

Norman Lloyd Johnson and Samuel Kotz. *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Wiley New York, 1977.

P Johnson and Montgomery Slatkin. Inference of microbial recombination rates from metagenomic data. *PLoS Genetics*, 5(10):e1000674, 2009.

Henry R Johnston and David J Cutler. Population demographic history can cause the appearance of recombination hotspots. *The American Journal of Human Genetics*, 90(5): 774–783, 2012.

John A Kamm, Jeffrey P Spence, Jeffrey Chan, and Yun S Song. An exact algorithm and efficient importance sampling for computing two-locus likelihoods under variable population size. *arXiv preprint arXiv:1510.06017*, 2015a.

John A Kamm, Jonathan Terhorst, and Yun S Song. Efficient computation of the joint sample frequency spectra for multiple populations. *arXiv preprint arXiv:1503.01133*, 2015b.

Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

J. F. C. Kingman. The coalescent. *Stoch. Process. Appl.*, 13:235–248, 1982a.

J. F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland Publishing Company, 1982b.

J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982c.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Jere Koskela, Paul A Jenkins, and Dario Spano. Computational inference beyond kingman's coalescent. *Journal of Applied Probability*, 52(2):519–537, 2015.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988.

Sergio Lukić and Jody Hey. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics*, 192(2):619–639, 2012.

G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. 160:1231–1241, 2002.

G.A.T. McVean, S.R. Myers, S. Hunt, P. Deloukas, D.R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304 (5670):581–584, 2004.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

P.A.P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71, 1958.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

S. Myers, R. Bowden, A. Tumian, R.E. Bontrop, C. Freeman, T.S. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

Simon Myers, Colin Freeman, Adam Auton, Peter Donnelly, and Gil McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature genetics*, 40(9):1124–1129, 2008.

Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104, 2012.

Rasmus Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 2000.

Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

Judea Pearl. Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, pages 133–136, 1982.

Andrzej Polanski and Marek Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, Sep 2003.

Raazesh Sainudiin, Kevin Thornton, Jennifer Harlow, James Booth, Michael Stillman, Ruriko Yoshida, Robert Griffiths, Gil McVean, and Peter Donnelly. Experiments with the site frequency spectrum. *Bulletin of mathematical biology*, 73(4):829–872, 2011.

Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

Jeffrey P Spence, John A Kamm, and Yun S Song. The site frequency spectrum for general coalescents. *arXiv preprint arXiv:1510.05631*, 2015.

Paul R Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10): 1680–1682, 2015.

Matthias Steinrücken, John A Kamm, and Yun S Song. Inference of complex population histories using whole-genome sequences from multiple populations. *bioRxiv preprint doi:10.1101/026591*, 2015.

M. Stephens and P. Donnelly. Inference in molecular population genetics. *J.R. Stat. Soc. Ser. B*, 62:605–655, 2000.

Fumio Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, 1983.

Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–595, 1989.

S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26:119–164, 1984.

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

John Wakeley and Jody Hey. Estimating ancestral population parameters. *Genetics*, 145 (3):847–855, 1997.

Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, Lewis C Becker, Ingo Ruczinski, Terri H Beaty, Sharon L R Kardia, Deborah A Meyers, Kathleen C Barnes, Diane M Becker, Nelson B Freimer, and John Novembre. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.*, 43:847–853, 2011.

Bruce Weir. *Genetic data analysis II: Methods for discrete population genetic data.* Sinauer Associates, Sunderland, MA, 1996.

M Ye, SV Nielsen, M Nicholson, YW Teh, P Jenkins, F Colchester, JWJ Anderson, and J Hein. Importance sampling under the coalescent with times and variable population. 2013. URL `https://www.stats.ox.ac.uk/__data/assets/pdf_file/0007/9889/Coalescent_Sampling_Report.pdf`.